On trees and forests

Meta-analysis and heterogeneity in practice

Joanna in ‘t Hout

On trees and forests

Meta-analysis and between-study heterogeneity in practice

Joanna in ‘t Hout

Cover design
Joanna in ’t Hout
Proefschriftmaken.nl || Uitgeverij BOXPress

Printed by
Proefschriftmaken.nl || Uitgeverij BOXPress

Published by
Uitgeverij BOXPress, ‘s-Hertogenbosch

The research presented in this thesis was performed at the Radboud Institute for Health Sciences, at the Department for Health Evidence, Radboud university medical center, Nijmegen, the Netherlands.

On trees and forests

Meta-analysis and between-study heterogeneity in practice

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus,
volgens besluit van het college van decanen
in het openbaar te verdedigen op dinsdag 12 januari 2016
om 16.30 uur precies

door

Johanna in 't Hout

geboren op 29 september 1966
te Borne

Promotor:	Prof. dr. J.J. Goeman
Manuscriptcommissie:	Prof. dr. G.P. Westert
	Prof. dr. K.G.M. Moons (Universiteit Utrecht)
	Prof. dr. T. Stijnen (Universiteit Leiden)

Contents

Published chapters

Chapter 2
IntHout J[a], Hooijmans CR[a], Ritskes-Hoitinga M, Rovers MM. Meta-Analyses of Animal Studies: An Introduction of a Valuable Instrument to Further Improve Healthcare. ILAR Journal. 2014;55(3):418-26.
[a] Equal contributions

Chapter 3
IntHout J, Ioannidis JP, Borm GF. Obtaining evidence by a single well-powered trial or several modestly powered trials. Statistical methods in medical research. Epub. October 14, 2012.

Chapter 4
IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC medical research methodology. 2014;14(1):25.

Chapter 5
IntHout J, Ioannidis JPA, Borm GF, Goeman JJ. Small studies are more heterogeneous than large ones: a meta-meta-analysis. Journal of Clinical Epidemiology. 2015;68(8):860-9.

Chapter 6
IntHout J, Ioannidis JPA, Rovers MM, Goeman JJ. A plea for routinely presenting prediction intervals in meta-analysis. (submitted).

Chapter 1

General introduction

General introduction

Information on almost any disease and cure is available on the internet, even for exotic ones. When the information is based on adequate summarization of high quality research, patient and doctor can communicate on a more equal level than they ever did before. However, much of the information that is so easy to grasp (and which may lead to much patient worry and anxiety), comes from sources that at best may be called dubious; the information may be motivated by commercial interests, or based on experiences of only a few patients instead of high quality clinical research. This was one of the main focuses of the evidence based medicine movement from 1991: identifying, critically appraising, and summarizing the available evidence.[1] In 1991 Sir Iain Chalmers founded the international Cochrane Collaboration with the objective "to give doctors, other health professionals and patients the evidence they need to make informed decisions about treatments".[2] Now, Cochrane Reviews are systematic reviews of primary research in human health care and health policy, that are internationally recognised as the highest standard in evidence-based health care.[3]

Good-quality information is available, especially for those who work at an institution that can afford to pay for subscriptions on scientific journals. By means of PubMed, Embase, Google Scholar or other search engines, results of peer reviewed medical research are accessible. However, the number of published trials is continuously increasing - in 2010 there were already 75 randomized controlled trials published per day and a plateau in growth was not yet reached.[4] Usually there are at least a few papers available on the same topic, each paper presenting the results of a different clinical study. If the results described by the different papers are similar this gives the impression that the conclusions in these papers are reliable and can be trusted. However, such similarity is often lacking. Results described in various papers may differ markedly, even for studies with similar designs and after correction for chance variation. This variability in results is called between-study heterogeneity. The reason for this heterogeneity is partly the presence of small differences in study set-up: setting (country, type of hospital), intervention (e.g. dosing), patient selection, timing and type of outcome measurements, etc. Also the quality of

the studies may fluctuate, with some studies meeting much higher quality standards than others. As a result, some studies may conclude that an intervention is highly effective whereas others may find that it is hardly better than, or as good as no intervention. Variation can also be caused by the application of different statistical approaches between studies. In practice selection and summarization of reliable high quality evidence is not so straightforward: the available evidence needs to be selected and summarized in a systematic and valid way, in "systematic" reviews. Selection of the evidence has resulted in guidelines to improve the completeness and quality of reporting by authors of primary studies, due to frustration over incomplete and misleading abstracts and study reports.[5]

Summarization is not only hampered by differences in reporting quality, but also by between-study heterogeneity. Heterogeneity can be present even if study selection criteria would always be carefully defined and applied. The reasons for heterogeneity can be clinical or statistical by nature. Clinical heterogeneity can be caused by differences in patient population, assessment times, assessment tools, medication regimen, etcetera. Statistical heterogeneity can be caused by differences in the statistical approaches of the studies. Some of them may have been adjusted for gender whereas others have been adjusted for co-medication and others may not have been adjusted at all. Some of them provide sufficient information to derive treatment effects and standard errors, whereas others not even present the number of patients included in the study. Analyses may be more or less advanced. These factors may result in higher levels of heterogeneity in the resulting meta-analysis than would be clinically expected.

Summarization of the evidence is done by means of meta-analysis: the statistical combination of the results of two or more primary studies. Traditionally the methodology for meta-analysis is based on large sample approximations[6], assuming that there are sufficient studies, that these studies are a random sample of all possible studies, and that heterogeneity is equal for all studies in the meta-analysis. Ideally, results of similar studies are combined to estimate the summary effect. However, sample sizes of the studies may differ substantially, e.g. because the research question of the meta-analysis was not always the primary question that the study aimed to answer. Large studies tend to provide more precise results than small studies. Therefore large studies should be more important than smaller studies in the meta-analysis, but how much more

important depends on the approach that is chosen for the meta-analysis. Two main approaches to meta-analysis are the fixed-effects and the random-effects approach. In a fixed-effects model it is assumed that all studies estimate the same intervention effect, even though in practice each clinical trial will result in a slightly different estimate for the treatment effect, because trials are performed with limited sample sizes. In the fixed-effects meta-analysis, the precision (inverse variance) of the estimated treatment effect defines the importance of the studies in the weighted summary: imprecise studies receive low weights and very precise studies receive high weights. Consequently, the summary estimate of the treatment is strongly influenced by the precise, large studies. In a random-effects approach it is assumed that the studies are not all estimating the same intervention effect, but estimate different, yet related intervention effects.[7] The efficacy of the intervention may vary, depending on differences in study set up. In the random-effects approach larger studies are more important than smaller studies, like in the fixed effects analysis, but now the weights are based on the inverse of the sum of the study imprecision plus the estimated heterogeneity. If the estimated heterogeneity is small, the summary effect estimate will be similar to the fixed effect estimate. However, if it is large, small studies will be almost as influential on the summary effect as large studies. Further, the standard errors of summary effects estimated with random-effects analyses tend to be larger than those of fixed-effects analyses.

The UK Cochrane Editorial Unit kindly provided us with the statistical data of systematic reviews of interventions in clinical studies included in the Cochrane Database of Systematic Reviews (CDSR) Issues of 2009-2013. We used these data to empirically evaluate various aspects of meta-analysis methodology and to design realistic simulation studies for more extensive evaluations. Most meta-analyses in the Cochrane Database of Systematic Reviews (CDSR) Issues 2009-2013 were based on only a few small studies: of the 3,263 selected meta-analyses, 1,025 (31%) were based on two studies and 1,226 (38%) on three to five studies. Besides most of the studies were small: overall, of the 20,185 primary trials 14,985 (74%) were small. And in more than 50% the between-study variation was estimated to be larger than zero.[8]

This thesis aims to give insight into the application of meta-analysis methodology and to reflect on the role of between-study heterogeneity in the realistic setting where most meta-analyses are based on just a few studies and where some of these studies are small or very small.

Outline of the thesis

Most medical systematic reviews including meta-analyses are dedicated to clinical (human) research. Fewer than 250 systematic reviews of preclinical animal studies had been published prior to 2010, as opposed to almost 6000 Cochrane Reviews of clinical studies to date.[9] However, some questions can only be answered by means of experimental animal studies. The number of meta-analyses of animal studies is increasing, but it is still very much lower than the number of meta-analyses of clinical studies, even though the statistical methodology is rather similar. To explain the methodology and to stimulate animal researchers to perform more meta-analyses, we have written an introductory tutorial, which is presented in **Chapter 2**. An interesting aspect of pre-clinical meta-analyses is that heterogeneity is often more than a "nuisance" parameter. Animal studies are usually designed to explore various treatments, dosages, and interventions, and are usually small. As with clinical questions, interest may be on the summary estimate of the intervention effect, but another important focus may be the reasons why results vary between settings. A large advantage of an animal study meta-analysis is that it may prevent the sacrifice of new animals if there are already relevant studies available. Another advantage is that meta-analyses may have a positive effect on the methodological quality of the primary animal studies in the long term, similar to the effect human meta-analyses had on clinical studies.[5]

In **Chapter 3** we investigate the basic question with regard to two possible approaches to find the best evidence on the effect of an intervention: is it preferable to conduct one large new trial or it is better to summarize existing trial results by means of a meta-analysis. The effects of three complicating factors are evaluated. First, the size of the existing trials: if a meta-analysis is preferred to a large trial, is this then also the case if only a few small trials are available for the meta-analysis? Second, the influence of reporting bias. It is well known that the papers that get published are not always representative of all studies that have been performed. Studies with positive findings tend to get published more often than those with negative findings, which results in an overrepresentation of papers with positive results. Clearly, if a meta-analysis is based on the available papers reporting bias might influence the conclusions: they tend to be too optimistic. The third issue is the between study variation, or

heterogeneity. What is the effect of heterogeneity on the result of the study or meta-analysis?

The DerSimonian and Laird (DL) approach[10], which is based on large sample assumptions, is widely used for random effects meta-analysis. However, as most meta-analyses only contain a few studies, this often results in inappropriate type I error rates. The method described by Hartung, Knapp[11-13], and by Sidik and Jonkman[14, 15] (HKSJ) is known to perform better when trials of similar size are combined. But evidence in realistic situations, where one trial might be much larger than the other trials, is lacking. In **Chapter 4** we assess the relative performance of the DL and HKSJ methods when studies of different sizes are combined. Therefore we simulated meta-analyses of 2–20 trials with varying sample sizes and between-study heterogeneity, and allowed trials to have various sizes, e.g. 25% of the trials being 10-times larger than the smaller trials. We compared the number of "positive" (statistically significant at $p < 0.05$) findings empirically using both approaches, using 689 meta-analyses from the Cochrane Database of Systematic Reviews (CDSR) Issues 2012. We also show how DL results can be converted to HKSJ results by means of a few steps.

Especially if there are both small and large primary studies in a meta-analysis, the size of the estimated heterogeneity is very important for the conclusions because of the weighting method in the random-effects approach. Results of small studies are often associated with "small study effects": the phenomenon that results of small studies tend to be more positive than those of larger studies.[16-18] This can be due to reporting bias but also to quality issues, both of which may occur more often in small trials. Possibly also the heterogeneity between small studies is different from the heterogeneity between large studies. In **Chapter 5** we evaluate the role of small studies in meta-analyses by exploring 3263 meta-analyses from the CDSR Issues 2009-2013. We focus specifically on the size of the heterogeneity in relation to trial size.

Heterogeneity is not only relevant for the weight of the studies in the meta-analysis: it also contains clinically relevant information. The existence of a positive heterogeneity estimate implies that there are differences in the intervention effect between the trials. This occurs often: in approximately half of the meta-analyses the estimates for the between-study variation are positive. It means that the treatment will appear more effective in some settings than in

others, which clearly is clinically relevant. However, most reviewers and readers are uncertain with respect to the clinical interpretation of the heterogeneity estimates. In **Chapter 6** we argue that the prediction interval is helpful in this context, because it shows the range of true treatment effects that is expected in future patients. Confidence intervals only estimate the mean treatment effect. In case of heterogeneity, prediction intervals will show a wider range of expected treatment effects than confidence intervals, and thus may lead to different conclusions. We investigated how often this occurred, and how often the conclusion was so different that it was indeed clinically relevant, using meta-analyses of the CDSR Issues 2009-2013.

Findings of these research questions, and perspectives for future research on meta-analysis methodology are presented in **Chapter 7.**

References

1. Montori VM, Guyatt GH. Progress in evidence-based medicine. JAMA. 2008;300(15):1814-6.
2. Champkin J. “We need the public to become better BS detectors”: Sir Iain Chalmers. Significance. 2014;11(3):25-30.
3. The Cochrane Collaboration [updated Downloaded on 30 July 2015]. Available from: http://www.cochrane.org/cochrane-reviews.
4. Bastian H, Glasziou P, Chalmers I. Seventy-Five Trials and Eleven Systematic Reviews a Day: How Will We Ever Keep Up? PLoS Med. 2010;7(9):e1000326.
5. Mullen PD, Ramirez G. The promise and pitfalls of systematic reviews. Annual review of public health. 2006;27:81-102.
6. Hoaglin DC. We know less than we should about methods of meta-analysis. Research Synthesis Methods. 2015.
7. Higgins JPT, Green S, Collaboration C. Cochrane handbook for systematic reviews of interventions: Wiley Online Library; 2008.
8. IntHout J, Ioannidis JPA, Borm GF, Goeman JJ. Small studies are more heterogeneous than large ones: a meta-meta-analysis. Journal of Clinical Epidemiology. 2015;68(8):860-9.
9. Ritskes-Hoitinga M, Leenaars M, Avey M, Rovers M, Scholten R. Systematic reviews of preclinical animal studies can make significant contributions to health care and more transparent translational medicine. Cochrane Database Syst Rev. 2014;3:ED000078.
10. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7(3):177-88.
11. Hartung J. An alternative method for meta-analysis. Biometrical Journal. 1999:901-16.
12. Hartung J, Knapp G. A refined method for the meta-analysis of controlled clinical trials with binary outcome. Statistics in Medicine. 2001;20(24):3875-89.
13. Hartung J, Knapp G. On tests of the overall treatment effect in meta-analysis with normally distributed responses. Statistics in Medicine. 2001;20(12):1771-82.
14. Sidik K, Jonkman JN. A simple confidence interval for meta-analysis. Statistics in Medicine. 2002;21(21):3153-9.

15. Sidik K, Jonkman JN. On constructing confidence intervals for a standardized mean difference in meta-analysis. Communications in Statistics-Simulation and Computation. 2003;32(4):1191-203.
16. Turner RM, Bird SM, Higgins JP. The impact of study size on meta-analyses: examination of underpowered studies in Cochrane reviews. PLoS ONE. 2013;8(3):e59202.
17. Pereira TV, Horwitz RI, Ioannidis JPA. Empirical evaluation of very large treatment effects of medical interventions. JAMA. 2012;308(16):1676-84.
18. Ioannidis JPA. Why most discovered true associations are inflated. Epidemiology. 2008;19(5):640-8.

Chapter 2

Meta-analyses of animal studies: an introduction of a valuable instrument to further improve healthcare

IntHout J[a] , Hooijmans CR[a], Ritskes-Hoitinga M, Rovers MM.
[a]Equal contributions. ILAR Journal. 2014;55(3):418-26

Abstract

In research aimed at improving human health care animal studies still play a crucial role, despite political and scientific efforts to reduce preclinical experimentation in laboratory animals. In animal studies, the results and their interpretation are not always straightforward, as no single study is executed perfectly in all steps. There are several possible sources of bias, and many animal studies are replicates of studies conducted previously. Use of meta-analysis to combine the results of studies may lead to more reliable conclusions and a reduction of unnecessary duplication of animal studies. In addition, due to the more explorative nature of animal studies as compared to clinical trials, meta-analyses of animal studies have greater potential in exploring possible sources of heterogeneity.

There is an abundance of literature on how to perform meta-analyses on clinical data. Animal studies, however, differ from clinical studies in some aspects, such as the diversity of animal species studied, experimental design, and study characteristics. In this paper we will discuss the main principles and practices for meta-analyses of experimental animal studies.

Introduction

Animal experimentation plays a vital role in research aimed at improving human health and health care. For example, in 2012 more than four million animal studies took place in Great Britain in a research context [1], and in a small country as the Netherlands almost 600,000 animal experiments were conducted [2]. Many of the animal studies are replicates of studies conducted previously. This is not a surprise as replication of study results is one of the main principles of science. However, how do we decide when we have enough (reliable) information about a specific topic for decision making? Meta-analysis of animal studies might be of use herein.

In general, meta-analysis is a tool to evaluate the efficacy of an intervention, using all available information. In addition, meta-analysis, especially cumulative meta-analysis, can help to visualize unnecessary duplication of animal studies [3, 4]. A cumulative meta-analysis is a series of meta-analyses in which each meta-analysis incorporates one additional study. When the meta-analyses are sorted chronologically, the display shows how the evidence accumulated, and how the conclusions have shifted over a period of time [5]. For example, a cumulative meta-analysis conducted by Sena et al. in 2010 on tissue plasminogen activator (rtPA) in stroke showed that the estimate of efficacy was already stable in 2001, after data from some 1500 animals had been reported. However, this meta-analysis was only conducted in 2010, and after 2001 another 1888 animals were used, which was not necessary to establish the effect of tPA for stroke [4]. Note that a number of these studies were not performed to establish the efficacy of rtPA but used rtPA as a positive control or as comparator for novel interventions. Nevertheless, meta-analyses of animal experiments are an important tool in reducing the amount of unnecessary animal studies.

There is an abundance of literature on how to perform meta-analyses on clinical data. Animal studies, however, differ from clinical studies in some aspects, for example, animal studies are much more diverse in their populations (e.g. species), design and study characteristics. In this paper we will discuss the main principles and practices for meta-analyses of experimental animal studies.

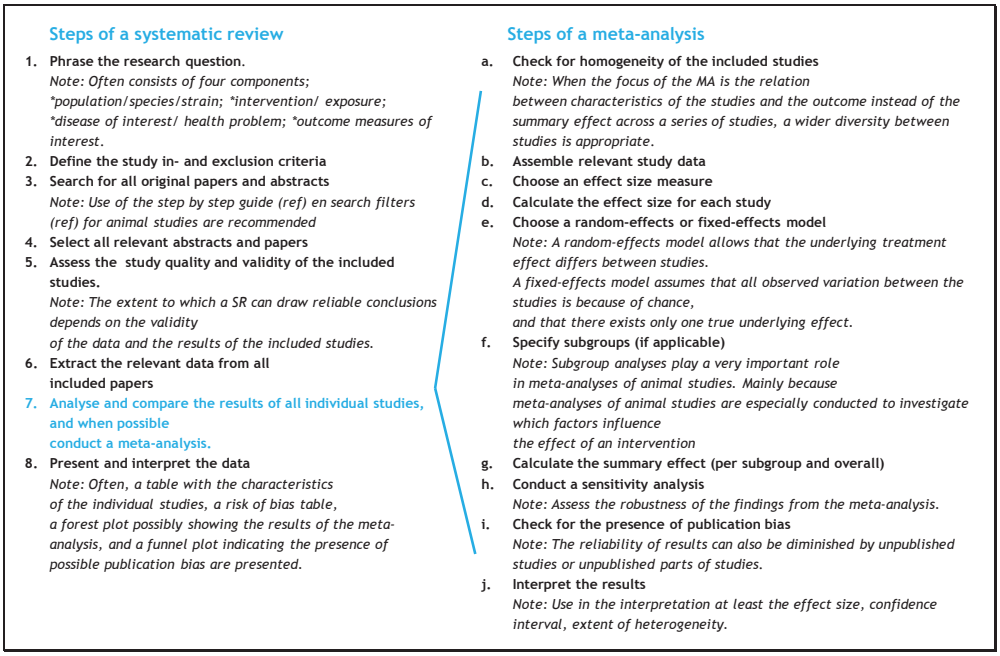

Figure 1. Steps to be taken in a systematic review (SR) and meta-analysis (MS) of animal studies. Figure 1 is party based on the general methods for Cochrane reviews [17].

Meta-analysis in the context of systematic reviews

When a scientist has an important research question to answer, there are often a variety of scientific approaches possible. One option is to design a new animal study. Another somewhat less common option is to conduct a systematic review of all animal studies. In a systematic review all research evidence relevant to a specific question is identified, appraised and synthesized in order to draw evidence based conclusions. In general, a systematic review results in a transparent overview of the available information, for example about the safety and efficacy of a treatment, and offers new information that was not available by analyzing each study individually. So, a systematic review might result in a better answer to the research question than a new animal experiment.

Systematic reviews are almost standard practice in clinical studies, but are not yet widely conducted in the field of laboratory animal science. Fewer than 250 systematic reviews of preclinical animal studies had been published prior to 2010, as opposed to almost 6000 Cochrane Reviews of clinical studies to date [6]. Given that many studies using laboratory animals aim at improving human health (and health care), it seems reasonable that research using animals be reviewed in a similar way and adhere to similarly high-quality standards. Some scientists even suggested that a more rigorous assessment of the results of animal studies in the form of a systematic review should be a prerequisite before starting studies in patients [7].

Eight different steps need to be taken when a systematic review is conducted (Figure 1). In one of these steps (step 7) the results of all individual studies are reported and compared, and when possible combined by means of a meta-analysis. This results in a quantitative summary of the knowledge that is available. However, a meta-analysis may also aim to assess the dispersion between the individual study effects. Although many systematic reviews contain meta-analyses, systematic reviews are also frequently published without a meta-analysis, especially when the included studies are too heterogeneous or seem to be seriously biased.

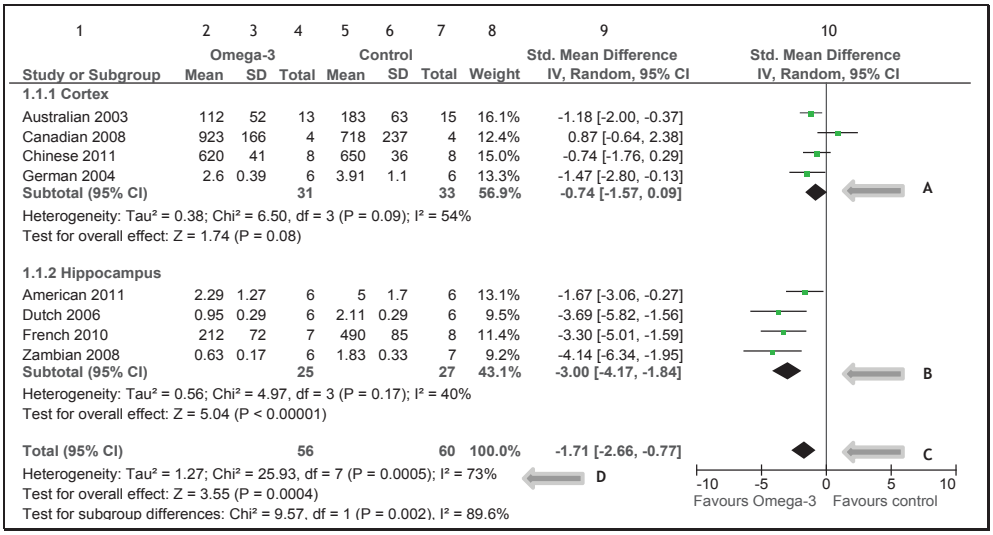

Figure 2. Forest plot, summarizing fictive results of eight individual studies, comparing the effects of omega-3 fatty acid supplementation vs. control treatment. Abbreviations: SD: Standard deviation; Std. Mean Difference: standardized mean difference; IV, Random: a random-effects meta-analysis is applied, with weights based on inverse variances; 95% CI: 95% confidence interval; df: degrees of freedom; Tau^2, I^2: heterogeneity statistics; Chi^2: the chi-squared test value; Z: Z-value for test of the overall effect; P: p value.

The results of a meta-analysis are displayed in a forest plot, see for example Figure 2. This plot allows readers to visualize and interpret the results of a meta-analysis. Figure 2 represents a forest plot summarizing fictive results of eight individual studies on the effects of omega-3 fatty acid supplementation on neuronal cell death in experimental Alzheimer's Disease (AD). Four studies assessed the amount of neuronal cell death in the cortex, the other four studies in the hippocampus. These are presented separately, as subgroups.

The first column shows the references of the included studies. Columns 2 through 7 show the raw data (mean, standard deviation (SD), and sample size (total)) of both the experimental omega-3 group and the control group concerning the amount of neuronal cell death due to omega-3 fatty acid supplementation in experimental AD. Neuronal death is measured using different scales, therefore the means of the various studies vary considerably (0.63 to 923). In this case some studies present cell death per $inch^2$ and others per mm^2.

Based on the raw data an effect estimate for each study can be calculated. In column 9, the study effects are represented as standardized mean differences (mean difference/ SD_{pooled}), so that the differences are expressed on a uniform scale; the fact that the scales of measurement varied across studies is no longer a problem. On the right, these differences, with their 95% confidence intervals (CIs), are presented with a central square and a horizontal line in the forest plot. The size of the central square is roughly equal to the size of the study, or more exactly, to the weight (column 8) that the study contributes to the combined effect [8]. The weight of a study varies with the statistical model used to pool the results (either fixed- or random-effects model). The vertical line represents the line of no effect.

In this example it was decided to combine the results of the individual studies. The resulting summary effects are depicted as black diamonds, per subgroup (arrows A and B) and for all studies combined (C). The location of the diamonds represent the point estimates (direction and size) of the treatment effect, and the width of the diamonds represent the 95% CI. In this fictive example, the diamond corresponding to the total effect (C) is located completely left of the vertical no-effect line. Therefore, we can conclude that there is a statistically significant reduction in the amount of neuronal cell death due to omega-3 fatty acid supplementation in experimental animal models for AD.

Last but not least, in this forest plot also the amount of heterogeneity is shown, expressed as Tau^2 (τ^2), together with a chi-squared test result (D). The τ^2 is an estimate of the between-study variation. The corresponding chi-squared test assesses whether the τ^2 is larger than zero, but is of limited importance as it is not very powerful when the number of studies is small and it gives no information on the extent of heterogeneity [9]. The I^2 (also at D) is a measure of inconsistency between the study results and quantifies the proportion of observed dispersion that is real, i.e. due to between-study differences and not due to random error [10]. It reflects the extent of overlap of the CIs of the study-effects. If I^2 is low (<25%), almost all observed variance is probably spurious. If the heterogeneity is large (e.g. I^2 >50%), we should speculate about reasons for the large "real" variance [9]. Animal studies are often quite exploratory and heterogeneous with respect to species, design, intervention protocols etc, compared to clinical trials. Exploring this heterogeneity is one of the added

values of meta-analyses of animal studies and might help to inform the design of a clinical trial.

Reasons to conduct a meta-analysis of animal studies

Although meta-analyses of animal studies are not yet routine in laboratory animal experimentation there are many advantages for doing so.

Results from a systematic review that includes a meta-analysis of animal studies may be more robust than results from single animal studies, if the meta-analysis is based on multiple high quality studies. Therefore the knowledge about the efficacy or side effects of a treatment or intervention may be more comprehensive. No single research endeavor is perfect, and experts are prone to bias. Combining studies that meet specific predefined criteria regarding content and quality may result in more reliable conclusions [11].

In many situations human evidence is lacking, for example in toxicity studies [12]. A critical evaluation of animal experiments, leading to information about the efficacy and possible side effects, can therefore inform clinical trial design and improve patient safety. For example, single animal studies are often too small to show whether or not a side effect is relevant or to present the full spectrum of side effects. When multiple small animal studies are combined, this will increase the power of the analysis and give more insight in the significance of a side effect. A meta-analysis about the effects of nimodipine (a calcium channel blocker) for acute stroke showed that there was no convincing evidence from the animal studies to substantiate the decision to start trials with nimodipine in large numbers of patients. However, at the time of the meta-analysis of the animal experiments, 29 clinical trials with circa 7500 patients were already conducted [13].

As mentioned above, the added value of meta-analyses of animal studies is the new knowledge that can be obtained by the evaluation of the heterogeneity between the studies. For example, a meta-analysis can make the impact of methodological quality on the effect size transparent. Bebarta et al. showed in 2003 that animal studies that do not utilize randomization or blinding are more likely to report a difference between study groups than studies that employ

these methods [14]. Part of the heterogeneity between animal studies can also be caused by differences of biological study characteristics (such as species, sex, age, dose, intervention schedule etc) on the main effect. The impact of such characteristics can be investigated by subgroup analyses or meta-regression (meta-analytical techniques to assess the relationship between study level covariates and effect size). A recent review about the effects of ischemic preconditioning (IPC) on ischemic reperfusion injury (IRI) in the animal kidney showed, for example, that the timing of the ischemic preconditioning greatly influenced the efficacy. The late window of protection (IPC more than 24h prior to IRI) appeared to be much more effective than the early window of protection. In addition, it was demonstrated that the IPC was more effective in rats than in mice [15]. These results obtained from the subgroup analyses resulted in the design of a new clinical trial focusing on the late window of protection instead of the early window which was used so far. This shows how meta-analyses may affect the design of future animal or clinical experiments.

So far, many papers and guidelines have been published regarding meta-analyses for clinical data [16-18]. Guidance has been published also for animal data, albeit not to the same extent [19]. Animal studies differ from clinical studies in some aspects, which has to be taken into account when performing a meta-analysis.

Differences between meta-analyses of animal and human studies

In human research the goal of a meta-analysis of clinical trials is generally to estimate the overall effect size of an intervention in order to aid decision making in clinical practice. In contrast, meta-analyses of animal studies are more exploratory and their results can be used to generate new hypotheses and guide the design of clinical trials. The purpose might be to summarize the effect of an intervention, to establish the relation between two variables, to summarize a parameter in a single group, or to evaluate heterogeneity between studies. The size of an effect, for example of an intervention effect in an animal model, is in itself not particularly useful information. This is partly because animal studies are so diverse in their populations (e.g. species), design and study characteristics; a pooled effect size is less meaningful compared to clinical trials.

However, as the studies in a meta-analysis are addressing a similar question, the direction of the effects is meaningful.

In addition, because animal studies offer a wider range of possibilities to examine toxicity of interventions or study pathology and mechanisms of disease than provided by clinical trials, meta-analyses of animal studies have a greater potential in exploring possible sources of heterogeneity compared to meta-analyses of clinical studies [19, 20]. We believe that one of the major added values of meta-analyses of animal studies is the insight that can be obtained by the evaluation of the heterogeneity between the studies.

The methods used for meta-analysis of animal studies are largely similar to clinical meta-analyses but in some aspects they are somewhat different [19]. For example, an animal study may contain both a placebo group and sham groups. In addition, animal studies are in general much smaller and more heterogeneous than clinical trials. Furthermore, the methodological quality of the included animal studies is often poor, which increases the risk of bias [21].

Meta-analysis of animal studies step by step

As mentioned before, a meta-analysis is often part of a systematic review. Once the review is started, i.e. once the scientific research question has been formulated, the objectives, study selection criteria, outcomes of interest and methodological approach should be described prospectively in a meta-analysis protocol. A protocol format for systematic reviews of animal intervention studies is submitted by de Vries et al. We recommend that authors register and/or publish the protocol, thereby allowing for feedback on the proposed methodology and insight in changes during the review process. When the studies have been selected and the relevant study results gathered, the statistical synthesis of the results – the meta-analysis – can be performed. This can be done for each outcome of interest, thus a systematic review can contain several meta-analyses per research question, for example, a meta-analysis for the outcome measure mortality, and one for the number of animals with weight increase. Each meta-analysis summarizes with statistical methods the results of those studies that reported on that outcome. A meta-analysis requires that at least two but preferably more studies are available.

We suggest the following key-elements and steps, amongst others inspired by the Cochrane Handbook [17], which all will be taken in a reproducible sequence when performing a meta-analysis of animal studies:

A. Check whether or not the included studies are homogenous enough to conduct a meta-analysis

An important feature of a systematic review – and thus also of the meta-analyses in the review – is that the systematic review often addresses a broader research question than was addressed by the primary studies. Consequently, the selected studies may show diversity in animal species, types of outcomes, measurement times etcetera. However, in order to be able to provide a meaningful answer to a research question like "what is the effect of this intervention on weight increase", a group of studies must be sufficiently homogeneous in terms of animals, interventions, designs and outcomes. Heterogeneity can be diminished by prospectively defining strict in- and exclusion criteria and making only sensible comparisons. It is therefore important to conduct a meta-analysis always in collaboration with an expert from the field.

On the other hand, if the aim of the meta-analysis was to determine factors (study characteristics) which influence the overall effect, especially the variation in the effect size is of interest and much more diverse studies may be included. In this case, where the focus of the MA is the relation between characteristics of the studies and the outcome, a wider diversity between studies is appropriate than when the focus is mainly on the summary effect across a series of studies.

B. Assemble the relevant study data

For each outcome of interest, for example weight change, data must be gathered for each study and treatment group, like columns 2-7 in the forest plot (Figure 2). Most of these can be extracted from the original publications. If data were presented only in graphs, they can be measured with digital ruler software. When the required data are missing the authors of the study should be contacted, which can take some time.

Study results may be expressed on different scales of measurement: counts (e.g., number of animals deceased), and mean values with standard

deviations (e.g., for weight increase) are most common. Preferably, data per group (counts or means and SDs and the total number of animals) are gathered for each study. If only medians and ranges or interquartile ranges are provided, these must be collected. If only summary effects are provided, e.g. in the form of Odds Ratios with a standard error or CI, these must be used. If no more detailed information is available, even only p-values and numbers of animals per group, or the direction of the effect size (positive or negative) can be useful [16] .

Also, the study design is important. It must be recorded what type of animals were used, timing of the measurements, details on the intervention, and other study characteristics that may be useful for the interpretation of the result. For example, if there are two control groups, including a sham group, or if control groups are shared by several experiments, this must be recorded. Animals receiving the same intervention are often group housed, altering the experimental unit into cage instead of individual animal. Furthermore, the animal experiments included in a meta-analysis are often not independent: control groups may be shared by two or more experimental studies.

C. Choose an effect size measure

Once the relevant study results are gathered, they may be used in the meta-analysis. If count data were gathered, you can choose whether you want to present the result of the meta-analysis in odds ratios, risk ratios or risk differences.

When the study results are continuous, like weight change, the choice is between ‘normal’ differences of the group means, standardized mean differences, and normalized mean differences. For example, when weight changes are measured in different species, the interpretation of an intervention effect of 6 g in a study with mice is completely different from the same effect in a study with beagles. In such situations standardized differences are a useful effect size measure, because they express the difference between the groups relative to the standard deviation. For instance, the weight changes of the mice might vary between -5 g and +12 g, and the SD is 4 g. An increase of 6 g corresponds thus to an increase of 1.5 SD. However, beagles weight on average 10 to 11 kg, and the individual

changes in weights will be much larger than for mice. If the SD of the weight changes of beagles is 500 g, the increase of 6 g corresponds to a minor change of 0.012 SD. The relevance of the effect is thus much better reflected by standardized differences than by the original differences. An additional advantage is that the scale in the forest plot automatically indicates the relevance of the summary result. This is also clearly shown in the forest plot (Figure 2). A normalized mean difference can be used when the score of a normal, untreated, unlesioned sham animal is known or can be inferred. One of the advantages of this method is that the absolute difference in means can be expressed as a proportion of the mean in the control group, which might be more easy to interpret [19]. For mean differences, standardized and normalized differences the same data must be extracted from each study: mean values, SDs and total number of animals per group.

When time to a certain event (e.g. death) is the topic of interest, survival data must be provided. See Vesterinen et al. [19] for details.

D. Calculate the effect size for each study / study subgroup

Once all the relevant data are collected and an effect size is chosen, study results must be prepared so that they can be used in the meta-analysis. In the most simple situation, each study provided separate data for both treatment groups; these data can be directly used to calculate effect sizes per study. However, from time to time data must be pre-processed, for example when median values and ranges or interquartile ranges are reported instead of means and SDs. If the data seem sufficiently normally distributed, medians and ranges can be used to construct means and SDs [22].

If results of some of the selected studies are not presented per group but combined as effect sizes, they can also be used in the meta-analysis. Take for example a set of five studies; three studies show weight increase data per group (mean, SD and total number of animals), and two studies only present the mean difference between the groups with a 95% CI. In this case, the result of each study must be transformed into a mean difference with corresponding standard error before it can be used in the meta-analysis.

In animal studies often the same control group is used for multiple

experimental groups. Sharing a control group makes the comparisons of the experimental groups dependent of each other. When such comparisons are presented as independent comparisons in a meta-analysis the animals in the control group will be counted twice and the comparisons will receive too much weight in the estimation of the summary effect. Therefore some adjustment must take place. A simple option is to diminish the number of animals in the shared control group by splitting the 'shared' group into two or more groups with smaller sample size [23]. For example, a study with two experimental groups sharing a control group with 12 animals results in two comparisons, each with six animals in the control group. For advice on other complicated situations, see the guidance of Vesterinen et al. [19].

E. Choose a random-effects or fixed-effects meta-analysis model

Random-effects and fixed-effects models are two statistical approaches which are used to combine the study results. They are based on different assumptions.

A fixed-effects model assumes that all observed variation between the studies is because of chance [24] , and that there exists only one true underlying effect. In other words: the variation between the study results is only because of variation in sample sizes. This assumption is reflected in the calculations of the study weights. Larger studies receive more weight.

A random-effects model allows that the underlying effect size differs between studies, thus an effect size can truly be larger or smaller, depending on the study characteristics. This heterogeneity is reflected by I^2 and was discussed above. The assumption that effect sizes truly differ is in general not implausible, because studies may have used different doses, routes of administration, animals or procedures, or there may be other, unknown differences. The random-effects model results in an "average" effect estimate, whereas the fixed-effects model results in an estimate of the one, true, underling effect [24, 25]. The confidence interval of a random-effects estimate will reflect that there is some possible variation in the true study effects besides chance alone, and therefore may be wider than that of a fixed-effects estimate. The two sources of variance are also taken into account in the assigned study weights.

F. Whether a fixed-effects or a random-effects model will be used must be decided before the meta-analysis is performed, and although one may be tempted to look at the level of I^2, the decision must be a priori [10] and based on substantive arguments, independent of the level or significance of I^2 [24]. Due to the nature of and diversity in animal studies, random-effects models may better reflect reality.

Once the meta-analysis is done, the interpretation of the results should be consistent with the model that was chosen. Take for example a fixed-effects meta-analysis comparing treatments A and B, that results in a mean difference of 1.75 and a 95% CI from 1.5 to 2. Here, 1.75 is the best estimate of the common treatment effect, and the CI reflects the uncertainty around this estimate. As zero is not in the CI, we can be quite sure that treatment A is superior to B. However, if the same numbers are the result of a random-effects meta-analysis, the interpretation is different. Now, we can be rather sure that **on average** treatment A is superior to B, but the true treatment effect may differ between settings. See [24, 25] for more information.

G. Specify subgroups, if applicable

Sometimes it is expected that the effect size varies across subsets of studies, for instance if there are variations between species or between dosages or administration routes of an intervention. In other cases we may observe heterogeneity in the results of the meta-analysis and want to find an explanation. In both situations subgroup analysis or meta-regression may give insight in the relation between study characteristics and the effect size. For example, our forest plot shows separate subgroups for studies which assessed the amount of neuronal cell death in the cortex and in the hippocampus. This stratified meta-analysis partitions the heterogeneity and shows that the estimated between-study variation is smaller in the subgroups: τ^2 in the subgroups is 0.38 and 0.56, whereas τ^2 in the pooled analysis is larger than the sum: 1.27. This suggests that there may be subgroup differences, which is confirmed by the test for subgroup differences.

Subgroup analyses play a very important role in meta-analyses of animal studies. This is to some extent due to the explorative character of animal studies, but also related to the aims of meta-analyses of animal studies. Many meta-analyses are especially conducted to investigate which factors

influence the effect size. If subgroups significantly differ in the effects, this may be an indication not to pool the overall results. It is however important to realize that the results of subgroup analyses can be misleading, especially if subgroups are not pre-specified. When subgroup analyses are not pre-specified, the risk on false positive findings, i.e. non-existing relations between effect size and study characteristics, increases. Further, subgroup analyses are often observational and not based on randomized comparisons [26]. In addition, subgroup analyses are often conducted on small numbers of studies which impairs the power of the analyses. Therefore, the results should be interpreted with caution [27]. Results of subgroups are hypothesis generating.

H. Calculate the summary effect, per subgroup and overall

In general, the summary effect size is based on the effect sizes and the weights of the individual studies. It can be calculated by hand, but there are also packages that will perform the calculations and provide forest plots, for example RevMan (www.ims.cochrane.org/revman/download), which is free software developed by the Cochrane Collaboration. CMA (Comprehensive meta-analysis (www.meta-analysis.com)) is not free, but offers simpler data entry and more options than RevMan. Stata (StataCorp, College Station, Texas) and R (http://www.R-project.org/) also provide meta-analysis packages. In case of more complicated designs, like multiple treatment groups sharing one control group, or studies with two control groups instead of one, it is advisable to consult a statistician.

I. The way the weights of the individual studies are calculated for the overall analyses, is dependent of the model (fixed-effects or random-effects) that was chosen (step E). In random-effects models, small studies get larger weights and are thus relatively more important than in fixed-effects models.

If heterogeneity is the main topic of interest, differences between subgroups are of special importance. Subgroup analyses or meta-regression of the outcome in relation to study characteristics (e.g. species) can give more insight in possible causes of heterogeneity, but should be conducted with caution, see point F.

If the results of the selected studies are considered too heterogeneous for pooling, a comparison between the number of studies with findings in one direction, and those with findings in the other direction (irrespective of whether or not the findings were significant) can be done with a sign test [16]. If pooling of the studies is considered completely inappropriate, no combined estimate can be provided.

J. Conduct a sensitivity analysis

A sensitivity analysis is conducted to assess the robustness of the findings from the meta- analysis. Assumptions underlying the initial meta-analysis can be challenged by performing another meta-analysis with different assumptions. If the results of both meta-analyses are similar, then they seem robust. For example, a scientist decided that the duration of an intervention might affect the results. Initially, a short intervention was defined as between zero and 45 minutes; in the sensitivity analyses the threshold was set at 30 minutes. What happens to the results if the threshold is changed? When the conclusions of a meta-analysis significantly change, this should be discussed.

Also the quality of the primary studies is crucial for the reliability of the meta-analysis results and can be assessed with a variety of tools [28, 29]. If some of the studies are suspected to present biased results, the meta-analysis can be performed once without those studies in order to investigate the robustness of the combined result.

K. Minimize publication bias

Reliability of results can also be diminished by unpublished studies or unpublished parts of studies. The risk of publication bias can be estimated by means of funnel plots [27]. Funnel plots are also provided by standard software for meta-analysis.

At the start of the systematic review, a serious attempt to gather all relevant study results must be made, in order to minimize possible reporting bias. Reporting bias, or publication bias, is the consequence of not all relevant data being available. In general, the decision to publish study results may depend on the direction of the study results, and negative studies, that are often relatively small in sample size are sometimes not published. In case of

publication bias, meta-analyses may overestimate the true effect size [30]. Publication bias is no less of an issue for animal studies than for human studies [31].

L. Interpretation of results

A meta-analysis will result in a summary or overall effect with a 95% CI and a p-value. In our forest plot the overall effect was an SMD of -1.71, with a 95% CI ranging from -2.66 to -0.77 (arrow C). The pooled estimate -1.71 is an average effect, and effects of the original studies will be spread out around this average effect. The summary effect of a meta-analysis is expressed as a number, for example an OR of 0.6, but in animal studies it is often wiser to focus on the direction of the effect than on the size itself. This is in large parts due to the unavoidable heterogeneity between animal studies (large variation in species, intervention protocols etc) and the explorative nature of animal studies compared to clinical research.

The confidence interval contains all likely effect sizes. If a 95% CI for an OR ranges from 0.4 to 0.9 this means that the true effect is most likely an OR between 0.4 and 0.9. If the 95% CI of an OR contains the value 1 or the 95% CI of mean difference contains the value 0, this means that the treatment groups are not statistically significantly different at a significance level of 5%. This is also reflected in a p-value above 0.05. If groups are not statistically significantly different, this does not necessarily mean that the treatment groups are similar; the only conclusion that can be drawn from the meta-analysis is that there is insufficient evidence to prove that the groups are different. Note that in general in case of multiple testing the significance level should decrease.

Interpretation of the results of subgroup analyses (steps F and G) is even more challenging. Subgroups in meta-analyses of animal studies are often very small and remain quite heterogeneous as multiple characteristics in animal studies vary. Results of subgroup analyses should therefore be used to generate rather than test hypotheses.

The forest plot resulted in a p-value of 0.0004 for the overall intervention effect. In general, a p-value below 0.05 means that the treatment groups are statistically significantly different. However, this does not necessarily

mean that the groups are different in a relevant way. It may be that the meta-analysis was based on many studies and thus had high power, whereas the effect size was only minor. In such a situation the meta-analysis will result in a p-value below 0.05, but the difference between the interventions may be irrelevant. Therefore not only the p-value but especially knowledge on the direction and size of the effect including the 95% CI are essential for the interpretation.

Another result of the meta-analysis is the extent of heterogeneity, presented with τ^2 and I^2, and tested with a chi-squared test. In this paper we focus on I^2, a reflection of the inconsistency between the effects estimated by the individual studies in a meta-analysis. It describes the percentage of total variation across studies that is due to heterogeneity rather than chance, and lies between 0% and 100%. A value of 0% indicates no observed heterogeneity, larger values show increasing heterogeneity [10]. When I^2 is 40%, for example, this means that 40% of the observed variation of the study effects is due to heterogeneity (τ^2) and 60% due to chance. An I^2 value near 100% means that the observed variation of the study effects is almost completely due to heterogeneity. The Cochrane handbook [26] states that an I^2 between 50- 90% might be interpreted as substantial heterogeneity. In this case it may be useful to evaluate whether some study characteristics may be the reason of this high heterogeneity, in order to prevent this in future animal studies. Note however that when a subgroup is found that seems to be the cause of the variation between the study results, this should be interpreted with caution.

Finally the strength of the conclusions from the meta-analysis also depends on the findings of the sensitivity analysis (H) and the evaluation of possible publication bias (I).

Conclusion

In this paper we address why meta-analyses of animal studies are a valuable addition for science aimed at improving health and healthcare. By conducting a meta-analysis of animal experiments, often new and very valuable information can be obtained from the already published animal experiments. In other words, the decision not to conduct a meta-analysis of animal studies may result in a waste of information, animals and financial resources.

The steps of a meta-analysis of animal studies are in general comparable to the steps taken in a clinical meta analysis, and software designed for clinical meta-analyses can be used for most of these steps. It is important that only sensible comparisons are made in order to reduce the risk on false positive findings and diminish heterogeneity. It is therefore important to conduct meta-analysis always in collaboration with an expert from the field.

Furthermore, it is very important that each scientist conducting a meta-analysis of animal studies realizes that the quality of a meta-analysis is also dependent of the quality of the primary studies. Especially if the purpose of the meta-analysis is to inform healthcare policy or practice the original research needs to be both applicable and of sufficient quality. On the other hand, if the primary studies appear to be biased, meta-analyses may provide the empirical evidence for the impact of the bias. This might stimulate the use of adequate experimental design in future animal studies. As we learned from systematic reviews in the clinical field [32], it is to be expected that the methodological quality of animal studies will increase as a consequence of conducting systematic reviews.

Briefly summarizing; meta-analyses of animal studies expand the knowledge resulting from animal experiments.

References

1. Winston R. Animal experiments deserve a place on drug labels. Nature medicine. 2013;19(10):1204.
2. NetherlandsFoodandConsumerProductSafetyAuthority. Zodoende 2012; Annual review of the Dutch Food and Consumer Product Safety Authority on animal experiments and laboratory animals. Utrecht: 2013.
3. Lau J, Antman EM, Jimenez-Silva J, Kupelnick B, Mosteller F, Chalmers TC. Cumulative meta-analysis of therapeutic trials for myocardial infarction. The New England journal of medicine. 1992;327(4):248-54.
4. Sena ES, Briscoe CL, Howells DW, Donnan GA, Sandercock PA, Macleod MR. Factors affecting the apparent efficacy and safety of tissue plasminogen activator in thrombotic occlusion models of stroke: systematic review and meta-analysis. J Cereb Blood Flow Metab. 2010;30(12):1905-13.
5. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Cumulative meta-analysis. Introduction to meta analysis: John Wiley and Sons, Ltd; 2009. p. 371-6.
6. Ritskes-Hoitinga M, Leenaars M, Avey M, Rovers M, Scholten R. Systematic reviews of preclinical animal studies can make significant contributions to health care and more transparent translational medicine. The Cochrane database of systematic reviews. 2014;3:ED000078.
7. Sandercock P, Roberts I. Systematic reviews of animal experiments. Lancet. 2002;360(9333):586.
8. Perera R, Heneghan C. ACP Journal Club. Interpreting meta-analyses in systematic reviews. Ann Intern Med. 2009;150(4):JC2-, JC-3.
9. Deeks J, JPT H, Altman D. Identifying and assessing heterogeneity. In: Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions. Chichester: John Wiley & Sons; 2008.
10. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003;327(7414):557-60.
11. Ioannidis JP, Lau J. Can quality of clinical trials and meta-analyses be quantified? Lancet. 1998;352(9128):590-1.
12. Peters JL, Sutton AJ, Jones DR, Rushton L, Abrams KR. A systematic review of systematic reviews and meta-analyses of animal experiments with guidelines for reporting. Journal of environmental science and health Part B, Pesticides, food contaminants, and agricultural wastes. 2006;41(7):1245-58.

13. Horn J, de Haan RJ, Vermeulen M, Luiten PG, Limburg M. Nimodipine in animal model experiments of focal cerebral ischemia: a systematic review. Stroke. 2001;32(10):2433-8.
14. Bebarta V, Luyten D, Heard K. Emergency medicine animal research: does use of randomization and blinding affect the results? Academic emergency medicine : official journal of the Society for Academic Emergency Medicine. 2003;10(6):684-7.
15. Wever KE, Menting TP, Rovers M, van der Vliet JA, Rongen GA, Masereeuw R, et al. Ischemic preconditioning in the animal kidney, a systematic review and meta-analysis. PLoS One. 2012;7(2):e32296.
16. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Meta-analysis methods based on direction and p-values. Introduction to Meta-Analysis. Chichester, UK: John Wiley & Sons; 2009.
17. Higgins J, Green S. Cochrane Handbook for Systematic Reviews of Interventions: John Wiley & Sons; 2008.
18. Nordmann AJ, Kasenda B, Briel M. Meta-analyses: what they can and cannot do. Swiss medical weekly. 2012;142:w13518.
19. Vesterinen HM, Sena ES, Egan KJ, Hirst TC, Churolov L, Currie GL, et al. Meta-analysis of data from animal studies: A practical guide. Journal of neuroscience methods. 2013;221C:92-102.
20. Mapstone J, Roberts I, Evans P. Fluid resuscitation strategies: a systematic review of animal trials. The Journal of trauma. 2003;55(3):571-89.
21. Kilkenny C, Parsons N, Kadyszewski E, Festing MF, Cuthill IC, Fry D, et al. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. PLoS One. 2009;4(11):e7824.
22. Hozo SP, Djulbegovic B, Hozo I. Estimating the mean and variance from the median, range, and the size of a sample. BMC medical research methodology. 2005;5:13.
23. Higgins JPT, Deeks J, Altman DG. Special topics in statistics. In: Higgins J, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions. Chinester, UK: John Wiley & Sons; 2008.
24. Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. BMJ. 2011;342:d549.
25. Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. Journal of the Royal Statistical Society Series A. 2009;172(1):137-59.

26. Deeks J, Higgins JPT, Altman DG. Analysing data and undertaking meta-analyses. In: Higgins J, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions. Chinester, UK: John Wiley & Sons; 2008.
27. Reade MC, Delaney A, Bailey MJ, Angus DC. Bench-to-bedside review: avoiding pitfalls in critical care meta-analysis-funnel plots, risk estimates, types of heterogeneity, baseline risk and the ecologic fallacy. Critical care. 2008;12(4):220.
28. Hooijmans CR, Rovers MM, de Vries RB, Leenaars M, Ritskes-Hoitinga M, Langendam MW. SYRCLE's risk of bias tool for animal studies. BMC medical research methodology. 2014;14:43.
29. Krauth D, Woodruff TJ, Bero L. Instruments for assessing risk of bias and other methodological criteria of published animal studies: a systematic review. Environmental health perspectives. 2013;121(9):985-92.
30. Sena ES, van der Worp HB, Bath PM, Howells DW, Macleod MR. Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. PLoS Biol. 2010;8(3):e1000344.
31. ter Riet G, Korevaar DA, Leenaars M, Sterk PJ, Van Noorden CJ, Bouter LM, et al. Publication bias in laboratory animal research: a survey on magnitude, drivers, consequences and potential solutions. PLoS One. 2012;7(9):e43404.
32. Mullen PD, Ramirez G. The promise and pitfalls of systematic reviews. Annual review of public health. 2006;27:81-102.

Chapter 3

Obtaining evidence by a single well-powered trial or several modestly powered trials

IntHout J, Ioannidis JP, Borm GF.
Statistical methods in medical research. Epub.Oct.2012.

Abstract

There is debate whether clinical trials with suboptimal power are justified and whether results from large studies are more reliable than the (combined) results of smaller trials. We quantified the error rates for evaluations based on single conventionally powered trials (80% or 90% power) versus evaluations based on the random-effects meta-analysis of series of smaller trials.

When a treatment was assumed to have no effect but heterogeneity was present, the error rates for a single trial were increased more than tenfold above the nominal rate, even for low heterogeneity. Conversely, for meta-analyses on series of trials, the error rates were correct.

When selective publication was present, the error rates were always increased, but they still tended to be lower for series of trials than for single trials.

We conclude that evidence of efficacy based on a series of (smaller) trials, may lower the error rates compared with using a single well-powered trial. Only when both heterogeneity and selective publication can be excluded, a single trial is able to provide conclusive evidence.

1. Introduction

There are conflicting opinions about whether clinical trials that are not sufficiently powered (i.e. with power below 80%) are justified. Some authors argue that the opportunity to perform such trials makes it easier to initiate and complete studies, while subsequent meta-analyses provide information similar to that from one large, well-powered study [1, 2]. Other authors object to this approach and stipulate that trials should have at least 80% or even 90% power for the outcome of interest [3, 4]. They argue that small studies make meta-analyses unreliable, mainly because of selective publication [5, 6]. In addition, if there is heterogeneity between the trials, meta-analysis is sometimes even avoided, leaving uncertainty about the interpretation of the results [7]. However, heterogeneity is also mentioned as one of the arguments to conduct several trials rather than a single large one, as the series of relatively smaller trials offer the opportunity to estimate the level of heterogeneity, which provides an indication of the generalizability of the results [8]. When heterogeneity is present, the effectiveness of a treatment will vary more than would be suggested by a single trial, so it may be premature or even erroneous to draw conclusions based on a single trial [4, 8, 9, 10, 11].

Although those issues have been discussed repeatedly, this has mainly been from a conceptual perspective. Here we try to provide more comprehensive evidence through a series of simulations. We aim to quantify the pros and cons of different approaches for the evaluation of treatments, such as an evaluation based on one single conventionally powered trial (80% or 90% power) versus an evaluation based on two trials, or on a series of smaller trials, each of which has modest power when seen in isolation (30% or 50% power). In particular, we evaluate the impact of heterogeneity and selective publication on the error rates and bias in the estimated effect sizes.

In sections 2 and 3 we present the extent of heterogeneity and selective publication that has been reported during the evaluation of treatments with a dichotomous health outcome. These estimates are used in our simulations. In sections 4 to 8 we evaluate the error rates and biases that may result when treatments are evaluated through single trials or through a series of trials. Finally, section 9 is devoted to the discussion and the conclusions.

2. Heterogeneity

The classical random effects model for a risk ratio meta-analysis assumes that trial results follow a distribution with:

$$E(D_i) = E(\log(RR_i)) = \delta_i, \quad \delta_i \sim \text{Normal}(\delta, \tau^2) \tag{1}$$

and

$$\text{var}(D_i \mid \delta_i) = \varepsilon_i^2 \text{ and } \text{var}(\delta_i) = \tau^2 \tag{2}$$

where $D_i = \log(RR_i)$ is the observed log-transformed risk ratio in trial *i*, and δ is the underlying overall mean of the log-transformed RR.

Usually the treatment effects of studies in a systematic review may differ. But if the study effects are more different from each other than one would expect due to random error (chance) alone, this additional between-trial variance is called heterogeneity. In order to use realistic values in our simulations for the heterogeneity we randomly selected 202 meta-analysis reports that compared two treatments by means of a dichotomous outcome variable from reviews of the Cochrane data base (25 August 2010). Our findings were similar to earlier reports: approximately half (51%) of the selected reviews had a point estimate for the between trial variance τ^2 equal to 0, in a similar range as previously reported percentages of 49% [12] and 37% [13].

In our selected reviews, 26% had an estimated τ^2 between 0 and 0.05, whereas 11% of the series had an estimated τ^2 between 0.05 and 0.15. The remaining 12% had heterogeneity up to 0.88. Based on these figures, we choose heterogeneity levels of 0, 0.05 and 0.15 for our simulations. As an illustration of a worst case scenario we also included heterogeneity of 0.8.

3. Selective Publication

Ross et al. investigated the publication rates for a cross section of trials that had been registered at ClinicalTrials.gov after December 31, 1999 [14]. They found that by June 2007, 61% of trials with a reported end date before 2004 had been published. For trials with an end date in 2004 or in 2005, 52% and 42% had been reported, respectively. The authors found no major difference between small studies with less than 160 participants and larger studies (43% and 46%, respectively). However, it is possible that non-registered trials may be more likely to remain unpublished and smaller trials may be less likely to be registered than large trials. Based on a review of trials supporting new drugs approved by the FDA, Lee et al. did report a relation between publication rates and sample sizes above 135 participants [15]. Nevertheless, a relationship between trial size and publication rates was not found in three other empirical studies summarized in a review by Hopewell et al. [16].

Trials with negative or null findings are clearly less likely to be published than trials with positive findings. Dwan et al. have reviewed a series of articles that have assessed publication rates of randomised controlled trials [17]. The percentages of published trials that the articles reported varied between 21% and 93%, with a median of 49%. The differences in publication rates between trials with and without a statistically significant outcome varied between 10% and 39%, with median 26%. Although some articles in the review were more recent than others, no strong trend was found. Recent articles that were not included in the review reported similar findings [15, 16, 18, 19].

4. Analysis method

We evaluated the error rates and bias that resulted from analyses of single trials and analyses of series of trials. We used the risk ratio (RR) as the metric of choice. When two or more studies were available for analysis, a random effects meta-analysis was carried out.

The commonly used method for random effects meta-analysis is the approach according to DerSimonian and Laird [20]. However, we used the method described by Sidik and Jonkman [21]. It is a simple approach that is similar to the

DerSimonian and Laird method, but it better preserves the error rates when the number of studies is small [21, 22, 23].

5. The error rates and bias of a series of trials: methods

Presence of between-trial variance and selective publication can lead to inflated error rates and bias in the results of the statistical analysis. In order to investigate its magnitude we simulated series of trials with a RR outcome under different degrees of heterogeneity and selective publication. The results of the subsequent analyses were used to evaluate the false positive error rates and the estimation bias.

False positive error rates and estimation bias

Selective publication 'favours' positive results, so it increases the false positive (FP) error rates. In addition, false positive error rates may often be more relevant than false negative rates (e.g. the consequences of adopting a non-effective treatment may be grave in terms of cost, potential toxicity, etc; while non-adopting a potentially effective treatment may not be as detrimental, especially when other effective treatments are also available) . Therefore, we focus on changes in the one-sided FP rates, calculated conditionally on H_0, i.e. the underlying average RR being 1:

$$\text{FP rate} = \frac{\text{No of series with positive analysis of published trials} \mid \text{RR=1}}{\text{No of simulated series}} \qquad (3)$$

Here, a positive analysis is defined as an analysis with a one-sided statistically significant result (P<0.025) in favour of the treatment under investigation (and for single trials also a more stringent criterion: P<0.0005). FP rates reflect the percentage of inefficacious treatments that result in a positive analysis.

In order to evaluate the bias, i.e. the effect of heterogeneity and selective publication on the final treatment estimate, we present the (geometric) mean of the estimated RR (MERR):

$$\text{MERR} = \exp(\text{mean of log risk ratios of all analyses of published trials}). \qquad (4)$$

In absence of bias, the MERR should be equal to the RR used to generate the simulation data. The bias is the difference between the MERR and the RR used to generate the simulation data (RR_{true}).

Trial and series characteristics

We simulated series of trials with an RR outcome under different degrees of heterogeneity and selective publication. Also the power, the risk in the control group and the number of trials in the series were varied. For each scenario, i.e. combination of these parameters we simulated 30,000 series of trials with RR_{true} =1 for the evaluation of the error rates, and 10,000 series with RR_{true}=0.5 and RR_{true} =0.8, corresponding to a large and modest treatment effect, respectively, for the evaluation of the estimation bias. Based on predefined publication rates, the simulation algorithm randomly selected which trials in a series were 'published', i.e. available for analysis, and which were not. On the available trials an analysis was performed.

A series of trials consisted of 1, 2, 3, 5, 10 or 20 trials. Between the series, the risk in the control group p_0 was varied from 0.1 to 0.9, in steps of 0.2. The control risks varied between the trials in a series, but remained within a range of $\pm\ p_0/2$ around the series risk p_0 (or around $1-p_0$ when p_0 was larger than 0.5); for example, with a series risk of 0.3, the trial control risk varied between 0.15 and 0.45, with 0.7, the trial control risk varied between 0.55 and 0.85. Trial sizes were based on 30%, 50%, 80%, or 90% power, the series control risk p_0 and an assumed risk ratio RR_{power} of either 0.5 or 0.8, respectively. The two-sided significance level used was 0.05. In the main simulations, each series consisted of trials with equal power.

Heterogeneity

For each combination of power, number of trials in the series, series risk p_0 and RR_{power} we generated a series of trials with a risk ratio RR_{true} equal to RR_{power}, i.e. 0.5 or 0.8 (alternative hypotheses) or RR_{true} equal to 1 (null hypothesis). Heterogeneity was superimposed as described in equation (1): the log-transformed risk ratios of each trial were normally distributed around δ, i.e. $\log(RR_{true})$, with τ^2 equal to 0, 0.05, 0.15 or 0.8.

Selective publication
The publication rates of trials with a positive statistically significant result (PRS) were set at 50%, 70%, 90%, or 100%. In a series, the publication rate of non-significant results (PRNS) was set to be 0% (no selective publication), 10%, 20%, or 30% smaller than the PRS for that series.

Series of trials with mixed power and power-related selective publication
Although there was no definitive published evidence for a relationship between trial size and selective publication (see section 3), there may be a relationship between the power of a trial and its probability of publication. Therefore, we executed a limited number of simulations for series with trials of mixed power where the publication rates were lower for the modestly powered trials. We simulated series of 5, 10 and 20 trials, and assumed that 80% of these trials were modestly powered (power 30%) and 20% conventionally powered (power 80%). The difference between publication rates for significant and non-significant findings δPR was set as follows: for trials with 30% power we assumed publication rates of 60% and 30%, and for trials with 80% power we assumed 90% and 80% publication rates for significant and non-significant results, respectively.

6. Error rates and bias of single trials when there is no selective publication

Table 1 and 2 show the FP rates for single trials when there is no selective publication. For each combination of between-trial variability τ^2 and power Table 1 presents the minimum and maximum error rate over all simulated control risks p_0 for a two-sided significance level $\alpha=0.05$ ($P<0.025$). Table 2 shows in a similar way the FP rates for a more stringent criterion: $\alpha=0.001$. Note that we evaluate the FP rates corresponding to the null hypothesis that the average treatment effect δ is 0 (corresponding to RR=1, equation (1)). However, in case of treatment heterogeneity, an average treatment effect of zero ($\delta=0$) allows that some of the trials have a truly non-null treatment effect. Trial *i* may have been conducted in a population or under conditions that led to a true effect δ_i. The standard analysis of trial *i* tests the within-trial null hypothesis $\delta_i=0$ based on the within trial variance e_i^2. But, in order to test a nonzero average effect for $\delta=E(\delta_i)$, not only the within trial variance e_i^2 but also the between-trial variance τ^2 should be taken into consideration.

Table 1 FP rates and RR estimates of single trials when there was no selective publication.

τ^2	Power (%)	FP rates (%)		RR estimates	
		RR_{power}=0.5 RR_{true} =1	RR_{power}=0.8 RR_{true} =1	RR_{power}=0.5 RR_{true}=0.5	RR_{power}=0.8 RR_{true} =0.8
0.00	30	2.7-3.2	2.6-2.7	0.49-0.50	0.80-0.80
	50	2.6-2.9	2.6-2.7	0.49-0.50	0.80-0.80
	80	2.6-2.7	2.6-2.7	0.50-0.50	0.80-0.80
	90	2.6-2.7	2.6-2.7	0.50-0.50	0.80-0.80
0.05	30	4.7-15.1	14.1-22.9	0.49-0.50	0.80-0.81
	50	6.3-18.3	19.7-27.2	0.50-0.51	0.80-0.81
	80	9.6-22.9	26.1-33.0	0.50-0.51	0.80-0.81
	90	11.6-24.5	28.7-35.1	0.50-0.50	0.80-0.81
0.15	30	8.6-24.8	23.9-32.3	0.50-0.51	0.81-0.81
	50	12.3-28.3	29.6-35.9	0.50-0.51	0.80-0.81
	80	18.3-32.3	35.0-39.9	0.50-0.51	0.80-0.81
	90	21.0-33.7	37.0-41.3	0.50-0.51	0.80-0.81
0.80	30	21.9-37.8	37.3-41.8	0.51-0.52	0.82-0.83
	50	27.0-39.9	40.6-43.9	0.51-0.52	0.81-0.83
	80	33.1-41.8	43.3-45.7	0.51-0.52	0.81-0.83
	90	35.0-42.6	44.3-46.4	0.51-0.51	0.81-0.82

FP: False-positive; RR: risk ratio.

FP rates: Minimum and maximum FP rates over all control risks p0; RR_{power}: the risk ratio used for the power calculation; RR_{true}: the risk ratio used in the simulations.

Table 2 FP rates (%) of single trials when there was no selective publication and a stringent significance level (α =0.001, power based on α = 0.05 and 0.001).

τ2	Power (%)	RRpower = 0.5 Powered on α=0.05	Powered on α=0.001	RRpower = 0.8 Powered on α=0.05	Powered on α=0.001
0.00	30	0.1-0.3	0.1-0.1	0.1-0.1	0.1-0.1
	50	0.1-0.2	0.1-0.1	0.1-0.1	0.1-0.1
	80	0.1-0.1	0.1-0.1	0.1-0.1	0.1-0.1
	90	0.1-0.1	0.1-0.1	0.1-0.1	0.1-0.1
0.05	30	0.3-4.6	1.6-11.1	3.7-11.1	14.5-23.1
	50	0.6-6.8	2.5-14.5	7.9-15.8	18.0-26.1
	80	1.6-11.1	4.6-18.2	14.7-23.1	22.6-30.3
	90	2.4-12.8	5.8-20.2	17.7-26.1	24.7-31.9
0.15	30	1.2-13.3	6.6-22.3	12.1-22.3	25.7-33.2
	50	2.7-17.2	9.4-25.8	18.7-27.2	29.0-35.4
	80	6.7-22.3	13.9-29.2	25.9-33.2	32.8-38.2
	90	9.1-24.1	15.9-30.8	28.7-35.4	34.4-39.2
0.80	30	9.9-30.0	22.9-36.6	29.2-36.6	38.8-42.7
	50	15.8-33.3	26.4-38.7	34.4-39.6	40.5-43.8
	80	23.1-36.6	30.6-40.7	39.0-42.7	42.3-44.8
	90	26.0-37.7	32.4-41.4	40.4-43.8	43.1-45.1

FP: False-positive; RR: risk ratio.

FP rates: Minimum and maximum FP rates over all control risks p0; RRpower: the risk ratio used for the power calculation, with RRtrue (the risk ratio used in the simulations) equal to 1.

As an illustration, the solid curve in Figure 1 shows the probability density function for trial result D_i under H_0: $\delta_i = 0$, transformed into a standard normal density function by division through the root of the within-trial variance, neglecting the heterogeneity. The dotted line shows a similar curve, but based on the total variance, including heterogeneity. It is clear that the part of the latter curve that exceeds 1.96 is much larger than 2.5%. When the analysis of a single trial is based on a within trial test $\delta_i=0$, it is straightforward to show that the FP rate of the overall test $\delta = 0$ is:

$$\text{FP rate} = 1-\Phi\left(\Phi^{-1}(1-\alpha/2)\sqrt{\varepsilon_i^2/(\tau^2+\varepsilon_i^2)}\right) \quad (6)$$

where α is the two-sided significance level and Φ is the standard normal cumulative distribution function [10, 11]. Results from this formula and the simulations for the risk ratio presented in Table 1 and 2 are approximately similar. Higher powered studies produce lower values for ε_i^2 and therefore, in combination with heterogeneity, lead to higher FP rates. As shown in Table 2, the FP rates when a study is powered for α =0.001 instead of 0.05 are even higher, especially when power is high. In general the FP rates powered for α =0.001 and a power of 50% are in the same range as FP rates powered for α =0.05 and a power of 90%. The high error rates illustrate that when heterogeneity is present, the result of a single trial is not suitable to draw conclusions on the overall mean treatment difference δ. As expected, Table 1 also shows there was no bias in the RR estimates, since no selective publication was present.

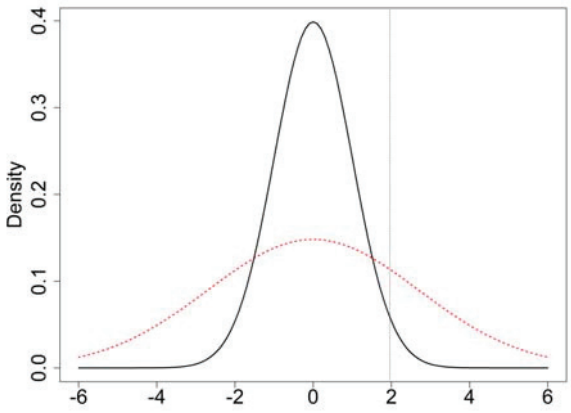

Figure 1. Probability density functions. Solid line based on trial-based (ε_i^2) variance, dotted line based on real ($\varepsilon_i^2+\tau^2$) variance needed to test the average treatment effect.

7. Error rates and bias

In current practice, the evaluation of a treatment can be based on a series of trials but also on one single study. This single study may be the only one that was conducted, but it might also be the only one out of a series of studies that was published. Take for example 100 series of two studies. With a publication rate of 80%, only 64 out of the 100 analyses will be based on both studies. In 32 cases, only one study will be available.

When there is no selective publication, the FP rates of series of trials were approximately 2.5% and there was no bias (results not shown).

FP rates and bias resulting from a setting with selective publication are presented in Table 3 and 4. When there was no heterogeneity, the RR estimates were biased, but FP rates were approximately 2.5% (Table 3). Table 4 shows FP rates and bias when there was heterogeneity. From some series, as a result of the selective publication, only one study was available for analysis and since the error rates of single studies are high (previous section), this resulted in markedly inflated error rates. Series of two trials, having a higher risk to result in only one available study than larger series, produced the highest error rates. When the number of trials in the series increased, the error rates decreased. The estimation bias was influenced by the selective publication and by the degree of heterogeneity.

Table 3 FP rates and RR estimates when there was selective publication but no heterogeneity.

Selective publication			N	FP rates (%)		RR estimates	
δPR (%)	PRS (%)	PRNS (%)		RR_{power}=0.5 RR_{true} =1	RR_{power}=0.8 RR_{true} =1	RR_{power}=0.5 RR_{true}=0.5	RR_{power}=0.8 RR_{true} =0.8
10	90	80	2	2.5-2.8	2.4-2.8	0.49-0.50	0.79-0.80
			10	2.5-3.1	2.5-2.8	0.49-0.50	0.79-0.80
	50	40	2	1.9-2.4	1.9-2.1	0.47-0.49	0.78-0.80
			10	2.4-2.9	2.4-2.7	0.48-0.49	0.79-0.80
20	90	70	2	2.5-3.0	2.4-2.9	0.48-0.50	0.79-0.80
			10	2.5-3.1	2.5-2.8	0.48-0.50	0.79-0.80
	50	30	2	2.0-2.6	2.0-2.2	0.44-0.49	0.77-0.79
			10	2.5-3.1	2.5-2.8	0.45-0.49	0.77-0.80
30	90	60	2	2.6-3.3	2.5-3.0	0.47-0.49	0.78-0.80
			10	2.5-3.1	2.5-2.8	0.47-0.49	0.78-0.80
	50	20	2	2.2-2.7	2.2-2.3	0.41-0.49	0.75-0.79
			10	3.2-3.9	3.2-3.5	0.42-0.49	0.75-0.79

FP: False-positive; RR: risk ratio; PRS: publication rate for significant results; PRNS: publication rate for not-significant results; N: number of conducted trials in each series. FP rates: Minimum and maximum FP rates over all control risks p0; RR estimates: mean estimated risk ratio; RR_{power}: the risk ratio used for the power calculation; RR_{true}: the risk ratio used in the simulations; δPR=PRS-PRNS.

Table 4 FP rates and RR estimates when there was selective publication and heterogeneity.

	Selective publication			τ^2 = 0.05				τ^2 = 0.80			
				FP rates (%)		RR estimates		FP rates (%)		RR estimates	
δPR (%)	PRS (%)	PRNS (%)	N	RR_{power}=0.5 RR_{true} =1	RR_{power}=0.8 RR_{true} =1	RR_{power}=0.5 RR_{true}=0.5	RR_{power}=0.8 RR_{true} =0.8	RR_{power}=0.5 RR_{true} =1	RR_{power}=0.8 RR_{true} =1	RR_{power}=0.5 RR_{true}=0.5	RR_{power}=0.8 RR_{true} =0.8
10	90	80	2	3.4-9.7	6.4-12.1	0.49-0.50	0.79-0.79	8.8-14.2	12.7-14.7	0.47-0.48	0.76-0.77
			5	2.6-3.3	2.9-3.5	0.49-0.50	0.79-0.79	3.1-3.4	3.2-3.5	0.48-0.48	0.77-0.77
			10	2.6-3.2	2.9-3.4	0.49-0.50	0.79-0.79	3.1-3.5	3.2-3.5	0.48-0.48	0.77-0.77
	70	60	2	3.5-13.8	8.7-18.8	0.48-0.50	0.79-0.79	12.6-22.5	19.8-23.7	0.47-0.48	0.75-0.77
			5	2.8-4.8	3.8-5.3	0.48-0.50	0.79-0.79	4.5-5.7	5.3-5.8	0.47-0.48	0.76-0.77
			10	2.7-3.4	3.0-3.5	0.49-0.50	0.79-0.79	3.2-3.7	3.5-3.7	0.47-0.48	0.76-0.76
	50	40	2	3.2-14.7	8.5-20.5	0.46-0.49	0.77-0.78	13.1-24.3	21.6-26.3	0.44-0.47	0.72-0.74
			5	3.0-9.2	6.2-11.5	0.47-0.49	0.78-0.78	8.5-13.0	11.7-13.3	0.45-0.46	0.72-0.74
			10	2.8-4.3	3.4-4.8	0.47-0.49	0.78-0.78	4.2-5.1	4.7-5.1	0.45-0.46	0.72-0.74
20	90	70	2	3.8-12.7	8.4-16.1	0.48-0.49	0.78-0.79	11.6-18.7	16.9-19.1	0.45-0.46	0.73-0.74
			5	2.7-4.1	3.4-4.4	0.48-0.49	0.78-0.78	3.8-4.6	4.3-4.7	0.45-0.46	0.73-0.74
			10	3.0-4.1	3.5-4.6	0.48-0.49	0.78-0.78	4.0-4.8	4.5-4.9	0.45-0.46	0.73-0.74
	70	50	2	3.8-16.9	10.2-22.5	0.46-0.49	0.76-0.78	15.0-26.5	23.4-27.8	0.43-0.44	0.69-0.71
			5	3.0-6.8	5.4-7.9	0.46-0.49	0.77-0.78	6.5-8.5	8.2-8.6	0.44-0.45	0.70-0.72
			10	2.7-4.7	3.6-5.3	0.47-0.49	0.77-0.78	4.3-5.6	5.3-5.6	0.44-0.45	0.70-0.71
	50	30	2	3.4-16.5	9.6-22.8	0.44-0.48	0.74-0.77	14.7-26.9	23.9-28.8	0.39-0.42	0.64-0.67
			5	3.9-13.7	9.0-16.8	0.45-0.48	0.75-0.77	12.6-18.6	17.1-18.8	0.40-0.42	0.66-0.67
			10	3.2-7.3	5.2-8.4	0.45-0.48	0.75-0.77	6.8-8.8	8.3-8.9	0.41-0.42	0.66-0.67

Table 4 (cont.) FP rates and RR estimates when there was selective publication and heterogeneity.

	Selective publication			$\tau^2 = 0.05$				$\tau^2 = 0.80$			
				FP rates (%)		RR estimates		FP rates (%)		RR estimates	
δPR (%)	PRS (%)	PRNS (%)	N	RR_{power}=0.5 RR_{true} =1	RR_{power}=0.8 RR_{true} =1	RR_{power}=0.5 RR_{true}=0.5	RR_{power}=0.8 RR_{true} =0.8	RR_{power}=0.5 RR_{true} =1	RR_{power}=0.8 RR_{true} =1	RR_{power}=0.5 RR_{true}=0.5	RR_{power}=0.8 RR_{true} =0.8
30	90	60	2	4.2-15.9	10.3-20.2	0.46-0.49	0.77-0.77	14.5-23.1	21.0-23.6	0.42-0.44	0.69-0.70
			5	3.0-5.6	4.3-6.2	0.46-0.49	0.77-0.78	5.1-6.4	5.9-6.4	0.42-0.44	0.69-0.70
			10	3.1-5.3	4.1-6.2	0.46-0.49	0.77-0.78	5.0-6.5	6.2-6.8	0.43-0.44	0.69-0.70
	70	40	2	4.1-19.5	11.7-25.8	0.44-0.48	0.74-0.76	17.2-30.0	26.7-31.4	0.39-0.41	0.64-0.66
			5	3.5-10.4	7.6-11.9	0.44-0.48	0.75-0.77	9.9-12.4	12.0-12.6	0.39-0.42	0.65-0.66
			10	2.9-6.5	4.7-7.7	0.44-0.48	0.75-0.77	6.0-8.3	7.8-8.5	0.40-0.42	0.65-0.66
	50	20	2	3.7-18.3	10.8-25.1	0.40-0.47	0.71-0.75	16.4-29.4	26.2-31.4	0.34-0.38	0.56-0.60
			5	5.3-19.7	13.3-23.8	0.41-0.47	0.72-0.75	18.5-25.5	24.6-25.9	0.34-0.38	0.57-0.59
			10	4.7-13.5	9.8-15.3	0.41-0.47	0.72-0.75	12.8-15.4	14.5-15.4	0.34-0.38	0.58-0.60

FP: False-positive; RR: risk ratio; PRS: publication rate for significant results; PRNS: publication rate for not-significant results;
N: number of conducted trials in each series.
FP rates: Minimum and maximum FP rates over all control risks p0; RR estimates: mean estimated risk ratio; RR_{power}: the risk ratio used for the power calculation; RR_{true}: the risk ratio used in the simulations; δPR=PRS-PRNS.

8. The error rates and bias when an 'at least two trials' approach is used

In the previous section we saw that false-positive error rates can be considerable when any statistically significant evidence (based on single studies or on series of more trials) is accepted as evidence of efficacy. Another policy would be to accept evidence of efficacy only if it is based on at least two available trials: an 'at least two trials' approach [24, 25]. If the literature search that precedes the planned meta-analysis would result in identifying only one trial, this policy leads to the conclusion that no sufficient evidence for efficacy is available, whatever the outcome of that trial, and one wants to wait for at least two trials to become available. To investigate the FP rates and bias from this 'at least two trials' policy, we simulated the situation that only analyses on at least two published studies could provide conclusive evidence of efficacy. In the absence of selective publication, the FP rates were approximately 2.5% and there was no bias (results not shown).

When selective publication was present, the error rates were increased and treatment effects were overestimated (too low estimates of the risk ratios), see Table 5. The impact of the selective publication mainly depended on the heterogeneity and the absolute difference δPR between the publication rates for significant and not-significant studies. Therefore, we used these to summarize the results. Each row in Table 5 corresponds to a certain degree of selective publication and heterogeneity, and shows the ranges of the error rates and the bias, taken over all values of p_0 and series of 2, 3, 5, 10 or 20 trials, each with 30%, 50%, 80% and 90% power.

The error rates and bias increased when τ^2 or δPR increased. Yet, when both the publication rates for significant and non-significant studies were high, the increase in error rates and bias was less extreme than when those rates were low.

When selective publication is present, FP rates will inevitably be inflated and treatment estimates will be biased. However, when δPR was 10% or less, the FP rates remained below 5% and the bias remained below 15% (Table 5). When the difference between PRS and PRNS increased beyond 10%, the error rates more than quadrupled, up to a maximum of approximately 16%. Nevertheless, even when δPR for the series was 60% (data not shown), they were below the error rates for single trials without selective publication (Tables 1 and 2). In the scenarios with high heterogeneity (τ^2 of 0.8) and selective publication (publication rates of 50% and 20% for the statistically significant and non-significant trials, respectively), the estimation bias increased up to 33%. Overall, a two-trial policy notably reduced the FP rates, whereas the estimation biases remained similar (Tables 3 -5).

For series that consisted of a mix of modestly powered (30% to 50%) and conventionally powered (80% or 90%) trials, the FP rates were in between those of series of trials with only modestly and only conventionally powered studies (results not shown).

Finally, yet another approach would be to have an ‘at least three trials’ criterion, i.e. to require that evidence of efficacy should be based on at least three published studies. We also investigated the consequences of this policy and found results that were largely similar to the ‘at least two trials’ policy.

Table 5 FP rates and RR estimates of a two-trial policy, when there was publication bias.

τ^2	Selective publication			FP rates (%)		RR estimates	
	δPR (%)	PRS (%)	PRNS (%)	RR_{power}=0.5 RR_{true} =1	RR_{power}=0.8 RR_{true} =1	RR_{power}=0.5 RR_{true}=0.5	RR_{power}=0.8 RR_{true} =0.8
0.00	10	90	80	1.6-3.3	1.5-2.8	0.49-0.50	0.79-0.80
		70	60	0.8-3.2	0.8-2.8	0.48-0.50	0.79-0.80
		50	40	0.4-3.2	0.4-2.8	0.46-0.50	0.78-0.80
	20	90	70	1.2-3.4	1.1-2.9	0.48-0.50	0.79-0.80
		70	50	0.6-3.5	0.5-3.1	0.47-0.50	0.78-0.80
		50	30	0.2-3.3	0.2-2.8	0.43-0.49	0.76-0.80
	30	90	60	0.9-3.6	0.8-3.0	0.46-0.49	0.78-0.80
		70	40	0.4-3.7	0.4-3.1	0.44-0.49	0.77-0.80
		50	20	0.1-3.2	0.1-2.8	0.41-0.49	0.75-0.79
0.05	10	90	80	1.7-3.5	1.7-3.7	0.49-0.50	0.79-0.79
		70	60	1.0-3.7	1.0-3.9	0.48-0.50	0.79-0.79
		50	40	0.4-4.0	0.4-4.5	0.46-0.49	0.77-0.78
	20	90	70	1.4-5.0	1.4-5.7	0.47-0.49	0.78-0.79
		70	50	0.6-5.5	0.8-6.4	0.46-0.49	0.76-0.78
		50	30	0.2-6.3	0.3-8.0	0.42-0.48	0.74-0.77
	30	90	60	1.1-7.0	1.3-8.4	0.46-0.49	0.76-0.78
		70	40	0.4-8.3	0.6-10.2	0.44-0.48	0.75-0.77
		50	20	0.1-10.8	0.3-14.8	0.40-0.47	0.71-0.75
0.15	10	90	80	1.7-3.8	1.8-3.9	0.48-0.49	0.78-0.79
		70	60	1.0-4.0	1.0-4.1	0.48-0.49	0.78-0.79
		50	40	0.4-4.3	0.5-4.6	0.46-0.48	0.76-0.77
	20	90	70	1.4-5.6	1.6-5.9	0.47-0.48	0.76-0.77
		70	50	0.8-6.4	0.9-6.8	0.46-0.48	0.74-0.76
		50	30	0.3-7.7	0.5-8.5	0.41-0.47	0.72-0.74
	30	90	60	1.1-8.3	1.5-8.8	0.45-0.48	0.74-0.76
		70	40	0.5-10.1	0.8-11.1	0.43-0.47	0.71-0.74
		50	20	0.2-14.0	0.4-16.0	0.38-0.45	0.67-0.71

Table 5 (cont.) FP rates and RR estimates of a two-trial policy, when there was publication bias.

τ^2	Selective publication			FP rates (%)		RR estimates	
	δPR (%)	PRS (%)	PRNS (%)	RR_{power}=0.5 RR_{true} =1	RR_{power}=0.8 RR_{true} =1	RR_{power}=0.5 RR_{true}=0.5	RR_{power}=0.8 RR_{true} =0.8
0.80	10	90	80	1.9-3.9	1.9-3.9	0.47-0.48	0.75-0.77
		70	60	1.1-4.2	1.1-4.3	0.46-0.48	0.75-0.77
		50	40	0.5-4.6	0.6-4.8	0.44-0.47	0.72-0.74
	20	90	70	1.6-6.0	1.9-6.1	0.45-0.46	0.72-0.74
		70	50	1.0-6.8	1.1-7.0	0.43-0.45	0.68-0.72
		50	30	0.5-8.6	0.5-8.8	0.39-0.42	0.64-0.67
	30	90	60	1.5-9.1	1.9-9.3	0.42-0.44	0.68-0.70
		70	40	0.8-11.0	1.1-11.5	0.39-0.42	0.63-0.66
		50	20	0.4-15.9	0.5-16.4	0.33-0.38	0.56-0.60

FP: False-positive; RR: risk ratio.

FP rates: Minimum and maximum FP rates over all control risks p0; RR estimates: mean estimated risk ratio; RRpower: the risk ratio used for the power calculation; RRtrue: the risk ratio used in the simulations; PRS: publication rate for significant results; PRNS: publication rate for not-significant results; δPR=PRS-PRNS.

9. Discussion and conclusion

Robustness of the results of a single trial

In general, proof of efficacy of a treatment can be based on the results of a single study, or on a series of studies. As we discussed before, proof based on a single study is often considered preferable over evidence from a series of trials. However, when heterogeneity is present and only one trial has been carried out, the FP rates may be very high, even if no selective publication is present, and even if the heterogeneity τ^2 is only 0.05 (Tables 1 and 2).

It has been suggested that a single well-planned large trial is the best way to evaluate the efficacy of a treatment [26-28]. Such a trial should have a detailed protocol and be executed according to the highest standards. Besides having sufficient power, the inclusion criteria, treatments and endpoints of the trial should also reflect clinical practice [29, 30]. From our findings, we conclude differently; in general, a single large trial may not be a suitable way to evaluate a treatment. This is in accordance with authors who showed that in some cases the results of large trials have been challenged and refuted over the course of time [2, 8, 31]. Our conclusion supports the FDA and EMA guidance that in therapeutic areas which are known for their variation in study results or in which seemingly convincing results could not be confirmed by subsequent studies, it would be better not to draw conclusions from a single study, even if it is large [24, 25]. A single trial, even if it is very significant, may only be sufficient when the results of previous trials in a specific therapeutic area were fairly robust and when variations in the treatment effects and patient populations were of minor importance [29, 30, 32]. However, even in this situation a single trial may have its drawbacks, because the trial may be biased by the way it is actually conducted or reported [6, 17, 33]. The latter may not be clear from the publication and may go unnoticed in one trial. In a series of studies, it is less likely that all studies will suffer from the same type of bias; consequently their composite picture may be more informative than the result of a single large trial.

Robustness of the results of systematic reviews

In the absence of selective publication the FP rates of systematic reviews are much lower than those of single trials. However, Table 3 shows that the error rates of systematic reviews may be very high when there is selective publication. Even if the difference between the PRS and the PRNS is only 10%, the FP rate

may exceed 10%. Clearly, the error rates may be considerable when single studies as well as series of trials are accepted as evidence of efficacy.

Robustness of systematic reviews with an 'at least two trials' policy

Another policy would be to accept evidence of efficacy only if it is based on at least two trials. If the literature search that precedes the planned meta-analysis or review, results in only one trial, this policy leads to the conclusion that no sufficient evidence for efficacy is available, whatever the outcome of that trial. When selective publication can be excluded, our results suggest that this approach leads to correct error rates and no bias. But even when selective publication was strong, the error rates of series of at least two published trials were lower than those of single trials without selective publication (Tables 5, 1 and 2, respectively). However, even for a series of at least two available trials, a difference of more than 10% between the publication rates for significant and not-significant trials resulted in a substantial increase in FP rates and bias.

Robustness of an 'at least three trials' policy

The robustness of the evidence provided by at least three published trials was similar (results not shown). However, adoption of a higher number of trials policy may still be preferable, because the trials are more likely to reflect local variability, and the effects of potential biases in any of the trials may be diluted, unless the biases are pervasive and affect many (or, even worst, all) trials. For example, this situation may apply when the entire clinical research agenda is designed by the same team or investigator and all trials are almost perfect replicas of each other.

Robustness of the evidence provided by a heterogeneity-sensitive-trial

An alternative approach may be to conduct a heterogeneity-sensitive-trial [8] i.e. a trial that produces realistic results that reflect variations in clinical practice. The trial procedures have to be sufficiently loose and the protocol must grant great freedom to the investigators [2, 30, 32]. In addition, the heterogeneity between the centres should be taken into account in the analysis; therefore the trial should be analyzed using a random treatment effects approach, as if it were a meta-analysis of a series of smaller trials, one in each centre.
Still, the heterogeneity in such a heterogeneity-sensitive trial may be limited, since all centres use the same protocol, follow the same procedures, share the same steering committee, etc. The effect of variations in clinical practice may

therefore be underestimated. Shrier provides an extensive discussion of the advantages and disadvantages of a heterogeneity-sensitive trial versus multiple smaller trials [8].

Prospectively defined trial series and meta-analysis

An approach similar to the heterogeneity-sensitive trial is to organize a series of individual trials with a pre-specified prospective meta-analysis [5], possibly under supervision of a Research Network. Examples can be found in the pre-planned meta-analyses by the FICSIT group [34] and the Carotid Stenting Trialists' Collaboration [35]. Prospectively planned trials series may be easier to organize and may even reflect variations in clinical practice more accurately than heterogeneity-sensitive trials [36]. For most new interventions, especially drug treatments, it is very common to have a large agenda of multiple trials being conducted anyhow [37]. Meta-analyses on the resulting data should preferably be planned a priori [37].

Poor and modestly powered trials

Here we have explored modestly powered trials (those with power of 30% to 50%). Modestly powered trials have the advantage that they are relatively small and therefore easier to organise. However, selective publication may be an issue. Some recent publications do not reveal a relationship between size and selective publication [14, 19], however, others do [15]. It may be prudent to refrain from conducting trials with minimal power. When a trial has at least 30% power, as in this paper, its size will be approximately a quarter of the size of a trial with 80% power. When the power is 50%, its size is half the size of a trial with 80% power. When the differences between trial sizes are of this magnitude, the magnitude of the selective publication may not be relevantly different.

Another issue that may play a role is the quality of the studies. Studies with low power are relatively smaller and the observation that smaller trials are sometimes of poorer quality is another potential reason to discourage them [1]. However, as the size of trials with 30% or 50% is not dramatically less than the size of trials with 80% power, the quality of the former should not necessarily be lower than that of trials with 80% power. In addition, the problem of low quality trials may become less serious in the future due to increasing regulations and codes for proper procedures and trial conduct (GCP, CONSORT [38], trial

registration). We agree with Schulz et al. that measures to improve quality are more worthwhile than putting emphasis on high power [1].

Limitations

Our study has some limitations. First of all, we restricted ourselves to the risk ratio metric. Although this limits the scope of our findings, the findings are relevant because systematic reviews with risk ratio as primary outcome are very common in phase III, when conclusive evidence is sought. Borm et al. provided results for a continuous outcome [10, 11]. Furthermore, our simulations are a simplification of the real situation. For example, we have not taken into consideration the effect of updating meta-analyses on the error rates and bias [39]. Neither have we considered 'opportunistic' strategies, for example when it is decided to change strategy and to do an unplanned second large trial, when the first one was not wholly decisive. In addition, we used a standard random effects model, assuming equal heterogeneity for small and large trials. However, heterogeneity might different between large and small trials, possibly due to the involvement of a higher number of study centres, better design or more rigid quality requirements of the larger trials. In that case, a standard random effects model might be too simple. Moreover, we focused on power and FP rates, but in some circumstances it is possible false-negatives may also be important to consider, e.g. if there is no other treatment available about a medical condition and thus falsely dismissing an effective treatment would be an unfortunate choice. Finally, we focused on a scenario where a trial is judged on a single efficacy outcome assessment, but sometimes many efficacy and harm outcomes may be important to consider as well.

Conclusions

When evidence of efficacy is based on at least two published trials, the error rates are substantially lower than for evidence based on a single large trial, even when selective publication is substantial. Therefore, the evaluation of a treatment should preferably be through a series of trials. For these trials, 30% power may be sufficient. Only when both heterogeneity and selective publication can be safely excluded, a single trial is able to provide conclusive evidence of efficacy of a treatment.

References

1. Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. Lancet 2005; 365(9467):1348-1353, doi:10.1016/S0140-6736(05)61034-3.
2. Edwards SJL, Lilford RJ, Braunholz D, Jackson J. Why "underpowered" trials are not necessarily unethical. Lancet 1997; 350:804-7, doi:10.1016/S0140-6736(97)02290-3.
3. Halpern SD, Karlawish JH, Berlin JA. The continuing unethical conduct of underpowered clinical trials. JAMA 2002; 288(3):358-362, doi:10.1001/jama.288.3.358.
4. Borm GF, Den Heijer M, Zielhuis GA. Publication bias was not a good reason to discourage trials with low power. J Clin Epidemiol 2009; 62:47e53, doi:10.1016/j.jclinepi.2008.02.017.
5. Rothstein, HR, Sutton, AJ and Borenstein, M. Publication Bias in Meta-Analysis, in Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments (eds H. R. Rothstein, A. J. Sutton and M. Borenstein), John Wiley & Sons, Ltd, Chichester, UK, 2006. doi:10.1002/0470870168.ch1.
6. Song F, Eastwood AJ, Gilbody S, Duley L, Sutton AJ. Publication and related biases. Health Technol Asses. 2000;4(10):1-115. NHS Centre for Reviews and Dissemination, University of York, York, UK.
7. Ioannidis JP, Patsopoulos NA, Rothstein HR. Reasons or excuses for avoiding meta-analysis in forest plots. BMJ 2008; 336(7658):1413-5, doi:10.1136/bmj.a117.
8. Shrier I, Platt RW, Steele, RJ. Mega-trials vs. meta-analysis: Precision vs. heterogeneity? Contemporary Clinical Trials 2007;28:324-328, doi:10.1016/j.cct.2006.11.007.
9. Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. JAMA 2005; 294(2):218-228, doi:10.1001/jama.294.2.218.
10. Borm GF, Lemmers O, Fransen J, Donders R. The evidence provided by a single trial is less reliable than its statistical analysis suggests. J Clin Epidemiol 2009; 62: 711-715, doi: 10.1016/j.jclinepi.2008.09.013.
11. Borm GF, Donders R. A treatment should be evaluated by small trials. Response to letter. J Clin Epidemiol 2009; 62: 886-889, doi:10.1016/j.jclinepi.2009.03.006.

12. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003; 327(7414):557-560, doi: 10.1136/bmj.327.7414.557.
13. Ioannidis JPA, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analyses. BMJ 2007; 335(7626): 914–916, doi: 10.1136/bmj.39343.408449.80.
14. Ross JS, Mulvey GK, Hines EM, Nissen SE, Krumholz HM. Trial Publication after Registration in ClinicalTrials.Gov: A Cross-Sectional Analysis PLoS Med 2009; 6(9): e1000144, doi:10.1371/journal.pmed.1000144.
15. Lee K, Bacchetti P, Sim I. Publication of Clinical Trials Supporting Successful New Drug Applications: A Literature Analysis. PLoS Med 2008, 5(9): e191, doi:10.1371/journal.pmed.0050191.
16. Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. Cochrane Database of Systematic Reviews 2009 Jan 21;(1):MR000006.
17. Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan A-W, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. PLoS ONE 2008; 3(8): e3081, doi:10.1371/journal.pone.0003081.
18. Song F, Parekh-Bhurke S, Hooper L, Loke YK, Ryder JJ, Sutton AJ, Hing CB, Harvey I. Extent of publication bias in different categories of research cohorts: a meta-analysis of empirical studies. BMC Med Res Methodol 2009, 9:79, doi:10.1186/1471-2288-9-79.
19. Rising K, Bacchetti P, Bero L. Reporting Bias in Drug Trials Submitted to the Food and Drug Administration: Review of Publication and Presentation. 2008 PLoS Med 5(11): e217, doi:10.1371/journal.pmed.0050217.
20. DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clinical Trials 1986;7:177-188.
21. Sidik K, Jonkman JN. A simple confidence interval for meta-analysis. Stat Med 2002; 21:3153-3159, doi: 10.1002/sim.1262.
22. Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. Stat Med 2001; 20:825-840.
23. Henmi M, Copas JB. Confidence intervals for random effects meta-analysis and robustness to publication bias. Stat Med 2010; 29: 2969–2983, doi: 10.1002/sim.4029
24. Guidance for industry. Providing evidence of effectiveness for human drug and biological products. FDA: Washington, 1998.

25. Points to consider on application with 1. meta-analyses; 2.one pivotal study. EMEA: London, 2001.
26. Egger M, Ebrahim S, Smith GD. Where now for meta-analysis? Int J Epidemiol 2002; 31:1-5.
27. LeLorier J, Grégoire G, Benhaddad A, Lapierre J, Derderian F. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. N Engl J Med 1997; 337(8):536-542.
28. LeLorier J, Gregoire G. Comparing results from meta-analyses vs large trials. JAMA 1998; 280(6):518-519.
29. Flather M, Delahunty N, Collinson J. Generalizing results of randomized trials to clinical practice: reliability and cautions. Clinical Trials 2006; 3:508-512, doi: 10.1177/1740774506073464.
30. Zwarenstein M, Oxman A. Why are so few randomized trials useful, and what can we do about it? J Clin Epidemiol 2006; 59:1125-26, doi: 10.1016/j.jclinepi.2006.05.010.
31. Ioannidis JP, Cappelleri JC, Lau J. Meta-analyses and large randomized, controlled trials. N Engl J Med 1998; 338(1):59-2.
32. Matthews JNS. Small trials: are they all bad? Stat Med 1995; 14:115-26.
33. Jüni P, Altman DG, Egger M. Assessing the quality of controlled clinical trials. BMJ 2001;323:42-6, doi: 10.1136/bmj.323.7303.42.
34. Province MA, Hadley EC, Hornbrook MC, Lipsitz LA, Miller JP, Mulrow CD, Ory MG, Sattin RW, Tinetti ME, Wolf SL. The effects of exercise on falls in elderly patients. A preplanned meta-analysis of the FICSIT trials. JAMA 1995; 273: 1341–1347.
35. Carotid Stenting Trialists' Collaboration. Short-term outcome after stenting versus endarterectomy for symptomatic carotid stenosis: a preplanned meta-analysis of individual patient data. Lancet 2010; 376 (9746): 1062-1073, doi: 10.1016/S0140-6736(10)61009-4.
36. Giffin RB, Woodcock J. Comparative Effectiveness Research: Who Will Do The Studies? Health Affairs 2010; 29 (11): 2075-2081, doi: 10.1377/hlthaff.2010.0669.
37. Ioannidis JPA, Karassa FB. The need to consider the wider agenda in systematic reviews and meta-analyses: breadth, timing, and depth of the evidence. BMJ 2010; 341:762-765, doi: 10.1136/bmj.c4875.
38. Schulz KF, Altman DG, Moher D, for the CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010; 340:c332, doi: 10.1136/bmj.c332.

39. Borm GF, Donders R. Updating meta-analyses leads to larger type I errors than publication bias. J Clin Epidemiol 2009; 62: 825-830, doi:10.1016/j.jclinepi.2008.08.010.

Chapter 4

The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method

IntHout J, Ioannidis JP, Borm GF.
BMC medical research methodology. 2014;14(1):25.

Abstract

Background

The DerSimonian and Laird approach (DL) is widely used for random effects meta-analysis, but this often results in inappropriate type I error rates. The method described by Hartung, Knapp, Sidik and Jonkman (HKSJ) is known to perform better when trials of similar size are combined. However evidence in realistic situations, where one trial might be much larger than the other trials, is lacking. We aimed to evaluate the relative performance of the DL and HKSJ methods when studies of different sizes are combined and to develop a simple method to convert DL results to HKSJ results.

Methods

We evaluated the performance of the HKSJ versus DL approach in simulated meta-analyses of 2–20 trials with varying sample sizes and between-study heterogeneity, and allowing trials to have various sizes, e.g. 25% of the trials being 10-times larger than the smaller trials. We also compared the number of "positive" (statistically significant at $p < 0.05$) findings using empirical data of recent meta-analyses with ≥ 3 studies of interventions from the Cochrane Database of Systematic Reviews.

Results

The simulations showed that the HKSJ method consistently resulted in more adequate error rates than the DL method. When the significance level was 5%, the HKSJ error rates at most doubled, whereas for DL they could be over 30%. DL, and, far less so, HKSJ had more inflated error rates when the combined studies had unequal sizes and between-study heterogeneity. The empirical data from 689 meta-analyses showed that 25.1% of the significant findings for the DL method were non-significant with the HKSJ method. DL results can be easily converted into HKSJ results.

Conclusions

Our simulations showed that the HKSJ method consistently results in more adequate error rates than the DL method, especially when the number of studies is small, and can easily be applied routinely in meta-analyses. Even with the HKSJ method, extra caution is needed when there are ≤ 5 studies of very unequal sizes.

Background

The commonly used method for a random effects meta-analysis is the DerSimonian and Laird approach (DL method) [1]. It is used by popular statistical programs for meta-analysis, such as Review Manager (RevMan [2]) and Comprehensive Meta-analysis [3]. However, it is well known that the method is suboptimal and may lead to too many statistically significant results when the number of studies is small and there is moderate or substantial heterogeneity [4-10]. If a treatment is inefficacious and testing is done at a significance level of 0.05, the error rate should be 5%, i.e. only one in 20 tests should result in a statistically significant result. For the DL method, the error rate can be substantially higher, unless the number of studies is large ($\gg 20$) and there is no or only minimal heterogeneity [4-10].

Given this deficiency, alternative methods for random effects meta-analysis have been proposed. In particular, the method described by Hartung and Knapp [4-6] and by Sidik and Jonkman [11,12] (HKSJ method) is claimed to be simple and robust [13]. Simulations have shown that the HKSJ method performs better than DL, especially when there is heterogeneity and the number of studies in the meta-analysis is small [4-14]. This means that for most meta-analyses the HKSJ method might be more appropriate than the conventional DL method. In a sample of 22453 meta-analyses, Davey et al. show that the number of studies in a meta-analysis is often relatively small, with a median of 3 studies (Q1-Q3: 2–6), and only 1% of meta-analyses containing 28 studies or more [15]. Some detectable heterogeneity is present in about half of meta-analyses of clinical studies [15-18].

Based on earlier results that showed that the results of a single large trial were unreliable [19], we hypothesized that the meta-analyses methods, including HKSJ, would perform less adequately when the meta-analysis is carried out on a mixture of very unequal-sized studies, e.g. one large and several small trials. Such a situation is not uncommon. In a random sample of 186 systematic reviews of the Cochrane Database [18] the ratio between large and small trial sizes ranged between 1 and 1650, with a median of 5 and an interquartile range from 3 to 10. Sixty per cent of the reviews contained no large trials, but 40% had one trial that was at least twice as large as the median trial size, 25% had one trial

that was at least five times larger, and 10% had one trial that was even 10 times larger.

Although several simulations have shown that the HKSJ method performs better than the DL method, the focus in these studies was not on a systematic evaluation of the effects of specific trial size mixtures in combination with low trial numbers. They either only reported the overall results of various mixtures combined or they studied only a limited number of combinations. In order to investigate the impact of unequal study sizes, we used simulations, mimicking such realistic conditions rather than situations where trials have implausibly similar sample sizes. We focused on meta-analyses with small numbers of studies (up to 20) with a dichotomous outcome (odds ratio, relative risk) or a continuous outcome. To mimic the variation in trial sizes, we explicitly varied the sample sizes of the trials within the simulated meta-analyses, varying from scenarios where all trials in a meta-analysis were of equal size, to scenarios with only one large trial, 10 times as large as the other trials, or one small trial, 10 times smaller than the other trials.

In order to complement the simulations, empirical data, based on recent meta-analyses - added or updated in 2012 - from the Cochrane Database of Systematic Reviews (CDSR) of interventions were used to assess the number of nominally statistically significant findings (with $p < 0.05$) of both methods in practice. This allows to examine whether inferences would be very different based on these two models.

Currently not all standard software packages like Review Manager provide an option to perform an HKSJ analysis, although the HKSJ method is computationally not complicated and the importance of suitable methods for meta-analyses with small numbers of trials is apparent. Version 3.0 of Comprehensive Meta-analysis [3] will contain the HKSJ method (personal communication by Julio Sánchez-Meca, September 2013). Also the R package metafor [20] and the metareg command in Stata [21] include the HKSJ method. However, not everybody will be acquainted with the use of R or Stata. Moreover, use of these packages is not straightforward when a post-hoc conversion is desired, i.e. when the results of a DL random effects analysis must be converted to the HKSJ approach. In order to fill this gap, we show step by step how the HKSJ analysis can be performed without the use of these packages, when the results of a common random effects

(DL) meta-analysis are available, e.g. from a systematic review. This conversion is applicable for continuous outcomes and for outcomes where metrics are log-transformed, like the risk ratio (RR), odds ratio (OR), hazard ratio (HR) or Poisson rate. This simple modification of the common random effects analysis will improve the summary results, and it can be done through some basic calculations or a few statements in Excel. An Excel file is available as Supplemental file (Figure S.3 in Appendix 4) and on the web. R code for the metafor package is provided in Appendix 3.

The simulations, the selection of empirical data and the statistical analysis are described in the Methods section. In the Results section the error rates for the DL and HKSJ methods for several realistic simulated scenarios are provided. For the Cochrane meta-analyses, we present the number of nominally statistically significant findings with the DL and HKSJ methods. The conversion of DL results into HKSJ results is illustrated, including examples from systematic reviews as presented in the Cochrane Library.

Methods

We used simulated data as well as empirical data of the Cochrane 2012 Issues to evaluate the DL and HKSJ approaches. The pooled effect estimate is equal for both approaches, but the methods differ with respect to the calculation of the confidence interval and the statistical test. For DL, these are based on the normal distribution, whereas for the HKSJ method, they are based on the t-distribution with the degrees of freedom equal to the number of trials minus one, and a weighted version of the DL standard error. Detailed statistical methods are presented in Appendix 1.

Methods – simulations

Our first aim was to investigate the error rates of the HKSJ meta-analysis method in comparison to the common (DL) method for various realistic scenarios, i.e. combinations of study sizes, study size mixtures and heterogeneity in series of just a few trials. Therefore we simulated series of trials with two up to 20 studies, where each series provided the data for one meta-analysis. First, we considered series that consisted of equally sized trials, each with two groups of 25, 50, 100, 250, 500 or 1000 subjects. Second, we looked into series of trials with different trial sizes, i.e. the percentage of large trials was 25%, 50% or 75%,

e.g. a series of one large trial and three small trials. Average group sizes were 100, 250, 500 or 1000 subjects, and the large trials had 10 times more subjects than the small trials. For example, a series of six small (normal) and two large trials, with an average group size of 100, has group sizes of 31 and 308 in the small and large trials, respectively. Third, we simulated extreme scenarios, in which a series had only small trials, except for one large one, or only large trials, except for one small one. Both continuous and dichotomous outcomes were evaluated. For continuous outcomes, a normally distributed overall mean difference between the group means was simulated. In the trials with a dichotomous outcome, the event rates in the groups varied between scenarios and ranged from 0.1 to 0.9, in steps of 0.2. The heterogeneity was superimposed and set at I2 = 0, 0.25, 0.50, 0.75 and 0.9. I2 represents the heterogeneity, i.e. the degree of inconsistency in the studies' results, in comparison to the total amount of variation [16,22]. The levels correspond to no, low, moderate, high and very high heterogeneity, respectively [16].

Our aim was to evaluate the error rate, i.e. the percentage of statistically significant meta-analyses when the overall mean treatment difference was zero. Hence we simulated series with an overall treatment difference equal to zero and performed on each series a DL [1] and an HKSJ [11] random effects meta-analysis. The two-sided significance level was 0.05. For each scenario, we simulated 10,000 series of trials. In the ideal situation, 5% of the 10,000 meta-analyses should have a statistically significant result when the significance level is 0.05. For the scenarios with the dichotomous outcome we determined the error rate when the OR was evaluated (logistic model) and when the RR was estimated. In these cases, meta-analysis was done on the logarithmic scale, and the error rates were determined for OR = 1 or RR = 1. More details can be found in Appendices 1 and 2.

Methods - empirical data from the 2012 Cochrane Database of Systematic Reviews

Cochrane Reviews are systematic reviews of primary research in human health care and health policy, and are internationally recognised as the highest standard in evidence-based health care [23]. The aim of the Cochrane collaboration is to provide accessible and credible evidence to guide decision making in medicine and public health. We were very fortunate that the UK Cochrane Editorial Unit

provided us with the statistical data added to the CDSR in 2012, which allowed us to assess the number of statistically significant results in real data.
Many Cochrane reviews include multiple meta-analyses. Many of those overlap or are based on correlated data. Usually, the first analysis is the primary analysis. Hence, we decided to use per review only the first meta-analysis that was based on at least three studies. In order to maximize the number of meta-analyses, we used both the first continuous and the first binary outcome meta-analysis, whenever possible. Thus some systematic reviews provided none, and some provided one or two meta-analyses for our research. We always performed a random effects meta-analysis, even when the authors originally performed a fixed-effects analysis. Details can be found in Appendix 1.

It is impossible to determine which of the Cochrane reviews compared treatments that truly had equal efficacy. It is thus unknown which of the statistically significant results were in fact false positive findings, so we could not determine the false positive error rate. Hence we decided to present the total number of significant findings of the DL and HKSJ methods instead of the error rates. This provides an indication of the impact a change from DL to HKSJ would have in practice.

Results

Error rates for continuous outcomes

The left side of Figure 1 shows the error rates for the DL method for the simulated mixtures of trial sizes. In general with unequal-sized trials, the type I error of DL was substantially inflated even with minimal heterogeneity, while with equal-sized trials minimal or modest heterogeneity did not inflate the type I error substantially. Figure 1A shows the error rates for a setting with studies of equal size, Figure 1B for one small trial, 1C for equal numbers of large and small trials, and Figure 1D for a setting with one large trial, 10 times as large as the other trials. The heterogeneity levels are I^2 = 0, 0.25, 0.5, 0.75 and 0.9, and the average study group sizes range between 25 and 1000. Vertical bars refer to the minimum and maximum error rates over the group sizes. The lines connect the means of these error rates. The error rates should all have been 5% (0.05), but for $I^2 \geq 0.25$, DL error rates were too large, even for series of 20 trials. For example, DL error rates for meta-analyses of five studies ranged between 5.7% for equally sized trials and 14.7% for mixtures of trial sizes (Table 1). In contrast,

the error rates were too low (about 3-4%) when the I^2 was 0. DL results for other, less extreme, mixtures of trial sizes were in between the results shown.
In Figure 1 on the right side results for the HKSJ approach are presented. For equal trial sizes, the error rates of the HKSJ method were very appropriate. When the series contained only one small trial, the HKSJ error rates were approximately correct if the series consisted of more than five studies (Figure 1B). For series containing fewer trials, the error rates were higher, but not as high as the respective DL values. They were also too high when the percentage of small trials increased (Figure 1C). When there was only one large trial, the HKSJ error rates sometimes almost doubled (Figure 1D). When there was no heterogeneity, HKSJ error rates were roughly 5%. As expected, the group sizes had no impact on the error rates.

Figure 1 shows that the HKSJ method always outperformed the common random effects DL method. The HKSJ error rate was usually roughly 5%. However, some mixtures of sizes, especially when there is only one large trial, lead to a doubling of the error rate to 10%. This occurred especially when heterogeneity was only moderate.

Error rates for risk ratio outcomes

The results of the simulations for studies with a risk ratio outcome were quite similar to the error rates for the continuous outcomes, but there was more variation in the error rates: they depended on the group sizes and the risks (from 0.1 to 0.9). For low heterogeneity (I2 = 0.25), the DL error rates ranged from 2.2% to 15.5%, whereas the HKSJ rates were slightly better: 2.8–10.6%. However for I2 =0.9 the DL rates ranged from 6.4% to 33.7%, compared to HKSJ rates of 2.7% to 10.2%. When there was no heterogeneity (I2 = 0), the DL error rates ranged between 0.9% and 4.3%, and the HKSJ rates between 2.1% and 6.9%. For odds ratios, the results were again quite similar. See Table 1 for a selection of results, and the Additional file 2: Figure S1 and Additional file 3: Figure S2.

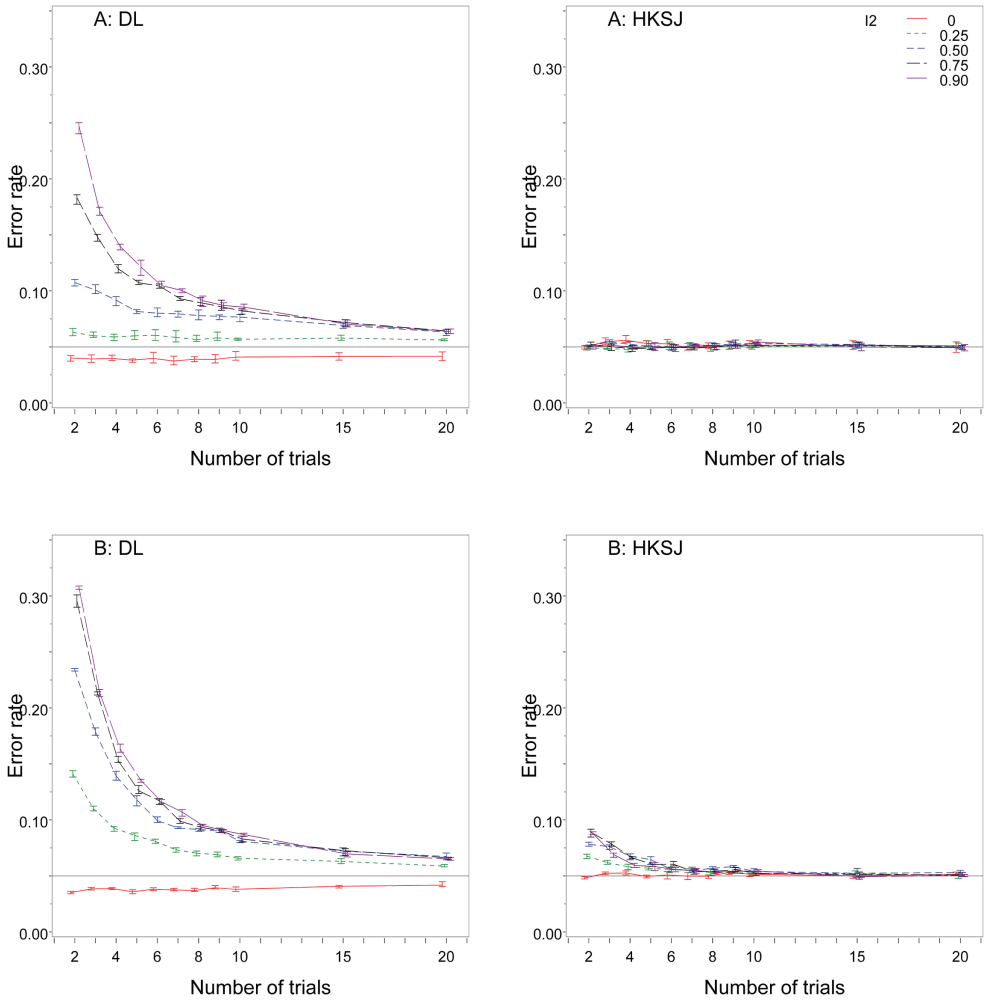

Figure 1 (A and B). DerSimonian-Laird and Hartung-Knapp-Sidik-Jonkman error rates for continuous outcomes, for various I^2 and mixtures of trial sizes. A: Equally sized trials; B: One small trial, 1/10th of other trials. Vertical bars refer to the minimum and maximum error rates over the group sizes. The lines connect the means of these error rates. DL: DerSimonian & Laird meta-analysis method. HKSJ: Hartung-Knapp-Sidik-Jonkman meta-analysis method.

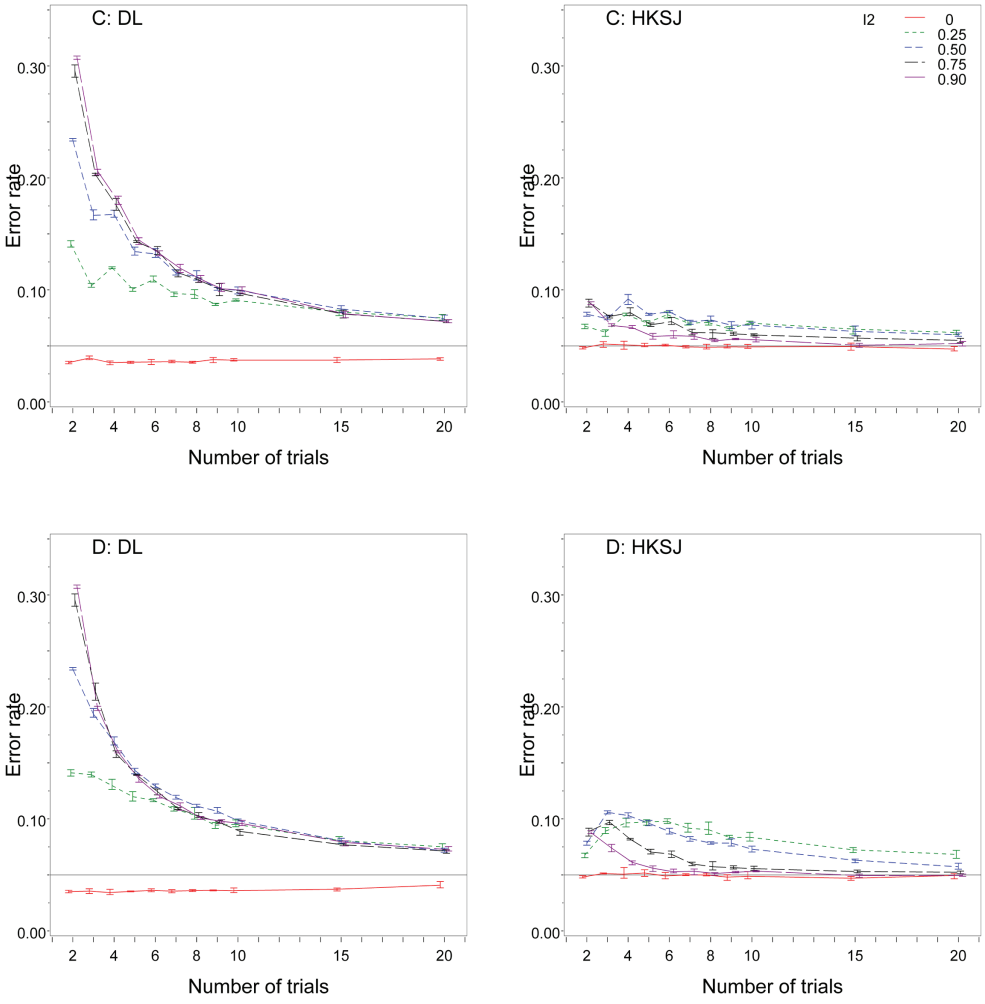

Figure 1 (C and D). DerSimonian-Laird and Hartung-Knapp-Sidik-Jonkman error rates for continuous outcomes, for various I^2 and mixtures of trial sizes. C: 50-50 small and large trials (ratio 1:10), D: One large trial (10 times larger than other trials). Vertical bars refer to the minimum and maximum error rates over the group sizes. The lines connect the means of these error rates. DL: DerSimonian & Laird meta-analysis method. HKSJ: Hartung-Knapp-Sidik-Jonkman meta-analysis method.

Table 1 Minimum and maximum error rates of DerSimonian-Laird and Hartung-Knapp-Sidik-Jonkman methods for mixtures of trial sizes

Outcome		No of	Equally sized		One small trial	
	I^2	trials	DL	HKSJ	DL	HKSJ
Continuous	0	2-20	3.4-4.6	4.5-6	3.4-4.5	4.7-5.4
0.25-0.9		2	6-25	4.7-5.4	13.8-30.9	6.5-9.2
		3	5.9-17.5	4.7-5.6	10.8-21.7	6-8
		4	5.6-14.2	4.5-5.5	9-16.8	5.6-7
		5	5.7-12.7	4.7-5.5	8.2-13.6	5.5-6.7
		10	5.6-8.8	4.8-5.6	6.4-8.8	5-5.6
		20	5.6-6.6	4.6-5.3	5.8-7.1	4.8-5.5
Risk ratio	0	2-20	0.9-4.2	2.1-6.9	2.8-4.1	2.7-6.5
0.25-0.9		2	2.5-26.3	2.8-6.7	14.3-33.7	6-10.2
		5	2.5-12.9	3.9-5.7	7.9-15	5.5-7.2
		10	2.6-8.9	2.7-5.4	6-9.7	3.8-5.7
Odds ratio	0	2-20	1.3-4.3	2.7-6.1	3-4	3-6.7
0.25-0.9		2	2.9-25.3	3.2-5.9	13.7-33.5	6.1-9.6
		5	3-12.7	3.9-5.3	7.9-14.4	5.4-6.9
		10	2.9-8.8	3.2-5.3	5.7-9.6	3.9-5.7
			50-50%		One large trial	
Continuous	0	2-20	3.3-4.1	4.6-5.4	3.2-4.4	4.5-5.7
0.25-0.9		2	13.8-30.9	6.5-9.2	13.8-30.9	6.5-9.2
		3	10.2-20.8	5.9-7.7	13.7-22.1	7.1-10.7
		4	11.9-18.4	6.6-9.6	12.6-17.3	5.9-10.5
		5	9.9-14.7	5.6-7.9	11.6-14.5	5.3-9.9
		10	9-10.3	5.4-7.2	8.5-10	5.3-8.8
		20	7.1-7.8	5-6.4	6.9-7.8	4.9-7.2
Risk ratio	0	2-20	3.0-4.1	2.7-6.8	2.7-4.3	2.7-5.5
0.25-0.9		2	14.3-33.7	6-10.2	14.3-33.7	6-10.2
		5	9.8-15.7	5.2-7.9	11.4-14.2	5.2-10.6
		10	8.6-11	4.8-9.1	7.3-10.1	3.6-8.7
Odds ratio	0	2-20	3-4	3-6.7	2.9-4.1	3-5.4
0.25-0.9		2	13.7-33.5	6.1-9.6	13.7-33.5	6.1-9.6
		5	9.9-15.8	5.3-8.1	11.6-14.2	5.2-10.5
		10	8.4-11.7	4.8-9.3	7.4-10.1	3.8-8.8

Error rates for the following scenarios: equally sized trials; one small trial, 1/10th of other trials; 50–50% small and large trials (ratio 1:10); one large trial (10 times larger than other trials). No of trials: number of trials. DL: DerSimonian & Laird meta-analysis method. HKSJ: Hartung-Knapp-Sidik-Jonkman meta-analysis method.

Empirical results for CDSR 2012

Selection of the first meta-analyses in the systematic reviews added in 2012 to the CDSR and based on at least three studies resulted in 689 meta-analyses (255 meta-analyses with a continuous outcome and 434 meta-analyses with a dichotomous outcome).

The continuous outcome meta-analyses were based on a median of five trials (Q1-Q3: 3–9) with a median ratio between the largest and the smallest trial of 5 (Q1-Q3: 3–10). Using the DL method, 130 (51.0%) of the 255 meta-analyses were nominally statistically significant compared to 102 (40.0%) when the HKSJ method was used (Table 2). Of the 130 meta-analyses that were significant with the DL method, 31 (23.8%) were not significant with the HKSJ method, while three meta-analyses were significant with the HKSJ method but not with the DL method. In the selection of meta-analyses based on at most five studies and with large ratios between the study sizes (ratio > 5) 13 (59.1%) of the 22 meta-analyses significant with the DL method were not significant with the HKSJ method and none of the meta-analyses was only significant with the HKSJ method.

The 434 dichotomous meta-analyses were based on a median of six trials (Q1-Q3: 4–10) with a median ratio between the largest and the smallest trial of 6 (Q1-Q3: 3–16). Of the 434 meta-analyses, 185 (42.6%) were nominally statistically significant with DL and 147 (33.9%) with HKSJ (Table 2). Of the 185 meta-analyses that were significant with the DL method, 48 (25.9%) were not significant with the HKSJ method, while the opposite scenario was seen in 10 cases. In the selection of small meta-analyses with large ratios between the study sizes 14 (50.0%) of the 28 meta-analyses significant with the DL method were not significant with the HKSJ method, while the opposite scenario occurred once.

Table 2 Number (%) of statistically significant Cochrane meta-analyses according to the DerSimonian-Laird and Hartung-Knapp-Sidik-Jonkman methods

Outcome	Selected meta-analyses	N	DL test significant	HKSJ test significant	HKSJ test not significant, positive DL test
Continuous	All	255	130 (51.0)	102 (40.0)	31/130 (23.8)
	Ratio > 5, ≤ 5 studies	46	22 (47.8)	13 (28.3)	13/22 (59.1)
Dichotomous	All	434	185 (42.6)	147 (33.9)	48/185 (25.9)
	Ratio > 5, ≤ 5 studies	76	28 (36.8)	15 (19.7)	14/28 (50.0)

All: all meta-analyses with a continuous or dichotomous outcome that fulfilled the following criteria: the first meta-analysis in a review in the Cochrane Database for Systematic Reviews Issues of 2012, based on at least three studies. Ratio >5, ≤ 5 studies: a selection of these meta-analyses based on at most five studies, where the ratio of the largest vs. the smallest trial size was > 5. DL: DerSimonian & Laird meta-analysis method. HKSJ: Hartung-Knapp-Sidik-Jonkman meta-analysis method. DL test significant: DL p-value <0.05; HKSJ test significant: HKSJ p-value < 0.05. Note that in a few cases the HKSJ test was significant when the DL test was not.

Summarizing, the DL method resulted in statistically significant results in 315/689 (45.7%) of the meta-analyses; 79 of these 315 “positive” DL results (25.1%) were not significant with the HKSJ method, while the opposite scenario (significant only by HKSJ) was rarely seen (14 meta-analyses). In the selection of small meta-analyses (≤ 5 studies) with large ratios between the study sizes (ratio > 5), the difference between the DL and HKSJ results was even larger.

Easy method for the conversion of DL into HKSJ results

We present two examples to illustrate how DL results can be used to carry out an HKSJ analysis, resulting in HKSJ-confidence intervals and p-values. An Excel file is available as Supplemental Figure S.3 in Appendix 4 and as web material. The results can also be created with R, Appendix 3.

Example 1: conversion to HKSJ for a continuous outcome

The first three columns of Table 3 show the results of a meta-analysis on the effect of zinc for the treatment of a common cold, published in a Cochrane review [24].The outcome was severity of cold symptoms scoring, and was based on a total of 513 participants. The first column shows the identifiers of the studies, the second column the results y_i of the individual studies and the third column contains the weights w_i from the DL analysis, copied from the review. Only these three columns are needed for the post-hoc calculations.

The following steps carry out an HKSJ analysis:

1. Determination of the standard error:
   a. Based on the overall summary difference y = −0.39, calculate the HKSJ factors
      wi×(yi-y)2 for each of the studies (see the fifth column for the results).
   b. Add the HKSJ factors and divide them by the sum of the weights. This results in 20.31/100 = 0.2031.
   c. Divide by k-1, whereby k is the number of studies. In this situation k = 5 and 0.2031/4 = 0.0508. This is the weighted variance of the pooled treatment effect according to the HKSJ approach.
   d. Taking the square root leads to the standard error: SE = √0.0508 = 0.225.

Table 3 Conversion of DerSimonian-Laird results into Hartung-Knapp-Sidik-Jonkman results for a continuous outcome: severity of cold symptoms

DerSimonian and Laird results			*Calculations for Hartung-Knapp- Sidik-Jonkman*	
Study	**Study results SMD** y_i	**Weights** w_i	$(y_i-y)^2$	$w_i \times (y_i-y)^2$
Kurugol 2006a	−0.04	24.0	*0.1225*	*2.94*
Kurugol 2007	−0.07	22.2	*0.1024*	*2.27*
Petrus 1998	−0.31	21.3	*0.0064*	*0.14*
Prasad 2000	−1.36	15.5	*0.9409*	*14.58*
Prasad 2008	−0.54	17.0	*0.0225*	*0.38*
	y = −0.39	*Sum: 100.0*		*Sum: 20.31*

5 studies, I^2 = 75.0%, τ^2 = 0.13

DL pooled result [95% CI]: SMD = −0.39 [−0.77, -0.02]; z = 2.05; P-value = 0.04

HKSJ pooled result [95% CI]: SMD = −0.39 [−1.02, 0.24]; t = 1.73; P-value =0.16 (df = 4)

SMD: Standardized mean difference. DL: DerSimonian & Laird meta-analysis method. HKSJ: Hartung-Knapp-Sidik-Jonkman meta-analysis method. CI: Confidence Interval, df: degrees of freedom, ×: multiplication sign. The pooled effect y and the weights w_i originate from the DL random-effects analysis.

2. Determination of the 95% confidence interval (CI):
   a. To determine the half-width of the 95% CI, the SE must be multiplied with the 97.5%-quantile of the t-distribution with k - 1 degrees of freedom. Its value can be obtained through Excel: TINV(0.05, k-1), where k is the number of studies. This results in 2.78, so the half-width of the 95% CI is 2.78*0.225 = 0.63. The t-value can also be found on the internet, for example at http://www.danielsoper.com/statcalc3/calc.aspx?id=10. The quantiles of the t-distribution can be found through statistical packages as well. In SPSS: select ‘compute variable’, function group ‘Inverse DF’, function IDF.T(.975,k-1), or in SAS: tinv(.975,k-1).
   b. The HKSJ 95% CI then is y ± half-width of the CI, i.e. -0.39 ± 0.63 or [-1.02; 0.24].

3. Determination of the p-value:
   a. Calculate the t-statistic: t = y/SE = −0.39/0.225 = −1.73. If the result is negative, as in this situation, simply change the sign, so t = 1.73.
   b. Determine the corresponding two-sided p-value with Excel: TDIST(1.73,4,2), or with the internet site http://www.danielsoper.com/statcalc3/calc.aspx?id=8. The two-sided P-value according to the HKSJ method then is 0.16.
   This p-value can also be obtained through SPSS: 'compute variable', function group 'CDF & noncentral CDF', function 'CDF.T'. This yields CDF.T(1.73, 4), similar to SAS, cdf('T', 1.73, 4) = 0.92066. The two-sided HKSJ p-value then is 2×(1–0.92066) ~ 0.16.

In this example on the efficacy of zinc, based on only five trials and high heterogeneity (I^2 = 75%), the results of the DL and HKSJ analyses differ substantially.

Example 2: conversion to HKSJ for outcomes that require a log transformation

When the outcome of the meta-analysis is a risk ratio (RR), odds ratio (OR), hazard ratio (HR) or Poisson rate, the analysis has to be conducted on the natural logarithm (ln) of the treatment effect. In all other aspects the procedure is exactly the same as for a continuous outcome. As an example we show the overall survival for post-remission therapy for adult acute lymphoblastic leukemia, comparing patients with and without a donor, as presented in a Cochrane Review [25]. The first three columns of Table 4 show the results of a meta-analysis with the HR as outcome.

Table 4 Conversion of DerSimonian-Laird results into Hartung-Knapp-Sidik-Jonkman results for a logarithm based outcome: Hazard Ratios

DerSimonian and Laird results			***Calculations for Hartung-Knapp-Sidik Jonkman***		
Study	Study results HR	Weights	$ln(y_i)$	$(ln(y_i)-ln(y))^2$	$w_i \times (ln(y_i)-ln(y))^2$
	y_i	w_i			
Cornelissen 2009	0.81	5.0	*−0.21*	*0.00*	*0.02*
De Witte 1994	0.67	2.1	*−0.40*	*0.06*	*0.13*
Fielding 2009	0.80	11.5	*−0.22*	*0.01*	*0.06*
Goldstone 2008	0.91	46.7	*−0.09*	*0.00*	*0.15*
Hunault 2004	0.56	2.9	*−0.58*	*0.18*	*0.53*
Labar 2004	0.98	9.3	*−0.02*	*0.02*	*0.16*
Ribera 2005	1.24	3.9	*0.22*	*0.13*	*0.52*
Sebban 1994	0.75	12.7	*−0.29*	*0.02*	*0.24*
Takeuchi 2002	0.95	3.9	*−0.05*	*0.01*	*0.04*
Ueda 1998	0.66	2.0	*−0.42*	*0.07*	*0.14*
	y = 0.86	*Sum: 100.0*			*Sum: 1.99*

10 studies, $I^2 = 0.0$, $\tau^2 = 0.0$.

DL pooled result [95% CI]: HR = 0.86 [0.77, 0.97]; z = −2.48; P-value = 0.013.

HKSJ pooled result [95% CI]: HR = 0.86 [0.77, 0.96]; t = −3.19; P-value =0.011 (df = 9).

HR: Hazard Ratio for donor versus no-donor; ln: natural logarithm; DL: DerSimonian & Laird meta-analysis method. HKSJ: Hartung-Knapp-Sidik-Jonkman meta-analysis method. CI: Confidence Interval, df: degrees of freedom, ×: multiplication sign. The pooled effect y and the weights w_i originate from the DL random-effects analysis on log scale.

1. Determination of the standard error:
   a. Calculate the natural logarithm of the pooled estimate: ln(y) = ln(0.86) = −0.15. Calculate the natural logarithms of the study outcomes (column 4) and use these to calculate the HKSJ factors wi×(ln(yi)-ln(y))2 for each of the studies (column 6).
   b. Add the HKSJ factors and divide them by the sum of the weights. This leads to 1.99/100 = 0.0199.
   c. As there are 10 studies, divide by k-1 = 9: 0.0199/9 = 0.0022.
   d. Taking the square root leads to the standard error: SE = √0.0022 = 0.047.

2. Determination of the 95% CI:
   a. On the ln scale, the half-width of the 95% CI is TINV(0.05, 9)×0.047 = 2.26×0.047 = 0.106 (Excel).
   b. The 95% CI for the ln HR is −0.15 ± 0.106, i.e. [−0.26; -0.04].
   c. The HKSJ 95% CI for the HR is [e-0.26; e-0.04], i.e. [0.77; 0.96].

3. Determination of the p-value:
   a. Calculate the t-statistic: t = ln(y)/SE = −0.15/0.047 = −3.19. Neglecting the negative sign, we obtain t = 3.19.
   b. Use Excel, Internet or a statistical package to calculate the two-sided p-value according to the HKSJ method, see Example 1. Excel: p-value = TDIST(3.19,9,2) = 0.011; SPSS: CDF.T(3.19, 9) = 0.995, so that the p-value is 2×(1–0.995) = 0.011.

In this example, results of the DL and HKSJ analyses hardly differ.

Discussion

The DL approach to random effects meta-analysis is still the standard method, almost to the exclusion of all other methods. This might be considered remarkable, bearing in mind the high false positive rates of the DL method which have been shown repeatedly with simulations [4-14] and also an empirical study suggesting that results are sensitive to the choice of random effects analysis method [26]. Thorlund et al. did an empirical assessment in 920 Cochrane primary outcome meta-analyses of >=3 studies of method-related discrepancies [26]. In total, 326 (35.4%) meta-analyses were statistically significant when the analysis was based on a t-distribution – as in the HKSJ method – and 414 (45%) when it was based on the normal distribution as in the DL method. Our evaluation of Cochrane meta-analyses of interventions resulted in a similar result: a substantially larger amount of significant findings with the DL method than with the HKSJ method. Our simulations suggest that among the DL significant findings in the Cochrane reviews there may be a considerable number of false positives.

DL results can easily be converted into HKSJ results, which have a much better performance. We confirmed this with simulations, for mixtures of trial size distributions in settings with up to 20 trials per meta-analysis. When there was heterogeneity, the mean error rates of the DL approach were consistently higher than those of the HKSJ approach, although also the latter doubled to 10% in scenarios with only one large trial. When there was no heterogeneity, the DL error rates were lower than 5%, and the HKSJ rates were approximately 5%.

However, there are some limitations with respect to the HKSJ analysis method. Although the error rates of the HKSJ method were closer to the 5% level than those of the DL method, our simulations showed that in some scenarios the HKSJ error rates more or less doubled, although the DL error rates could be more than four times too high in these same settings. Hence, the results of the HKSJ analysis are also not perfect. Like we hypothesized, the error rates were maximal if one of the trials in the meta-analysis was substantially larger than the other ones.

Further, when study numbers are small, the distribution of the treatment effects is unknown and does not necessarily follow the normal or t-distribution. Kontopantelis and Reeves [27] showed that with slight heterogeneity the coverage of the HKSJ method was consistently 94% when the true effects were not distributed according to the normal or t-distribution, but with larger heterogeneity the non-parametric permutation (PE) method of Follmann and Proschan [7] performed better than the HKSJ method. However, the PE method can only be performed when the number of studies is larger than five, whereas many meta-analyses are smaller [15]. Several other methods have been developed, like the Quantile Approximation (QA) method [28], the Profile Likelihood approach [29], natural weighting instead of empirically based weighting of studies [30], use of fixed effects estimates with a random effects approach to heterogeneity [31] and more recently, higher-order likelihood inference methods [32]. However, most of these methods are based on asymptotic statistics and they may therefore be less robust in case of a limited number of trials, or they remain difficult to use in practice, because no statistical packages are available to perform them and it is very difficult to carry out the calculations with standard software. Regarding the non-asymptotic, computationally straightforward QA method, Sánchez-Meca and Marín-Martínez [13] have already shown that it was outperformed by the HKSJ method. It would

require a very extensive evaluation to investigate the performance of all of these methods. We restricted ourselves to the HKSJ method, because of its computational simplicity and we show that HKSJ results can easily be derived from DL results.

As far as we know, we are the first to present systematically the error rates in relation to explicit trial size mixtures when the numbers of trials range from 2 to 20. Follmann and Proschan [7] show that for certain trial size mixtures and low numbers of trials the DL error rates can be highly increased, however, they did not evaluate the HKSJ method. The results reported by Hartung, Knapp and Makambi [4-6,8,9] imply that for meta-analyses of three, six or twelve studies the DL error rates for studies with similar sizes were closer to 5% than for studies of different sizes, and that the HKSJ method performed much better than DL in the latter situation. However they did not report the explicit relationship between the trial size mixtures and error rates as we do (Table 1). Sánchez-Meca and Marín-Martínez [13] also varied the sample size ratios in their simulations. They concluded that the average sample size scarcely affected the performance of the different methods, but this was based on the combined results of 5–100 studies and they presented no results of particular trial size mixtures.

As all studies show that in settings with few studies the HKSJ method always resulted in error rates closer to 5% than the DL method, the latter method should not be used and the HKSJ method should be the standard approach. To facilitate its more widespread application, the conversion of DL results into HKSJ results is presented step by step. At the same time, we urge caution when any random effects model, including HKSJ, is applied to situations where there are very few studies, and even more so when the sample sizes of the combined studies are very different. Even the HKSJ confidence intervals may be conservatively narrow in these situations and inferences may be spurious, if the confidence intervals are taken at face value.

Conclusions

Our simulations showed that the HKSJ method for random effects meta-analysis consistently results in more adequate error rates than the common DL method, especially when the number of studies is small. The HKSJ method can easily be applied routinely in meta-analyses. However, even with the HKSJ method, extra caution is needed when there are ≤5 studies of very unequal sizes.

Appendices

Appendix 1: Statistical details

Random effects meta-analysis model

For *k* studies, let the random variable y_i be the effect size estimate from the i^{th} study. The random effect model can be defined as follows:

$$y_i = \delta_i + e_i$$

for $i = 1, \ldots, k$, where $\delta_i = \delta + d_i$; e_i and d_i independent, $e_i \sim N(0, \in_i^2)$ and $d_i \sim N(0, \tau^2)$. $\in_i^2$ is the within-study variance, describing the extent of estimation error of δ_i, and the parameter τ^2 represents the heterogeneity of the effect size between the studies.

For studies with dichotomous outcomes where no events were observed in one or both arms, the computation of the random effects model yields a computational error. In these cases, before performing any meta-analysis, we added 0.5 to all cells of such a study.

Random effects analysis

Let w_i be the fixed effects weights, i.e. the inverse of the within-study variance $\hat{\in}_i^2$, and let $\hat{y}_F$ be the fixed effects estimate of δ.

Let *Q* be the heterogeneity statistic $Q = \sum w_i (y_i - \hat{y}_F)^2$. Then

$$\hat{\tau}^2 = \max\left(0, \frac{Q - (k-1)}{\sum w_i - \sum w_i^2 / \sum w_i}\right)$$

is an estimate of the variance τ^2.

The random effects estimate for the average effect size δ is

$$\hat{y}_R = \frac{\sum w_i^{\tau} y_i}{\sum w_i^{\tau}}$$

where

$$w_i^{\tau} = \frac{1}{\hat{\in}_i^2 + \hat{\tau}^2}.$$

The DerSimonian and Laird method estimates the variance of $\hat{y}_r$ by

$$\mathrm{var}_{DL} = \frac{1}{\sum w_i^r}$$

and uses the normal distribution to derive P-values and confidence intervals. In contrast, the Hartung, Knapp, Sidik and Jonkman method estimates the variance of $\hat{y}_r$ by

$$\mathrm{var}_{HKSJ} = \frac{\sum w_i^r \left(y_i - \hat{y}_r\right)^2}{(k-1)\sum w_i^r}$$

and uses the t-distribution with k-1 degrees of freedom to derive P-values and confidence intervals, with *k* the number of studies in the meta-analysis.

Heterogeneity estimates

Although $\hat{\tau}^2$ or *Q* can be used as measures of the heterogeneity, Higgins and Thompson [16] propose

$$I^2 = \frac{Q-(k-1)}{Q}$$

I^2 is a relative measure. It compares the variation due to heterogeneity (τ^2) to the total amount of variation in a 'typical' study ($\tau^2 + \epsilon^2$), where ϵ is the standard error of a typical study of the review [33]:

$$I^2 = \frac{\tau^2}{\tau^2 + \varepsilon^2} \qquad (1)$$

Appendix 2: The simulations

The parameters in the scenarios for the simulations

- the number of trials per series k = 2 – 20;
- the average group size in a series of trials: 25, 50, 100, 250, 500 or 1000 subjects per group per trial;
- the trial size mixtures: we simulated series with 25, 50 or 75% large trials, series with exactly one large or one small trial, and series where all trials were of equal size;
- the ratio of the study sizes: for the series with small and large studies, the large study was 10 times the size of a small study.

The simulations were programmed in SAS, version 9.2. The scenarios were evaluated 10,000 times, for heterogeneity levels I^2 = 0, 0.25, 0.5, 0.75, and 0.9, and at a nominal significance level α = 0.05 (two-sided).

A. The simulation for normally distributed outcomes

1. For each scenario, and each value of I^2, we used eq. (1) to calculate the variance τ^2. So

   $$\tau^2 = \varepsilon^2 \frac{I^2}{1-I^2} \qquad (2)$$

   where $\varepsilon^2 = \frac{1}{k} \sum \frac{2\sigma^2}{n_i}$, with n_i the groupsize of trial i (i = 1…k) and σ the standard deviation of the outcome variable of the trials. As σ is only a scaling factor and the results only depend on the ratio τ/σ, we have set σ =1 in the simulations.
2. For each trial i:
   a. We determined the 'true' trial effect size δ_i, where δ_i was a random draw from the normal distribution with mean 0 and variance τ^2.
   b. We generated the trial outcome based on a normal distribution with mean δ_i and variance $2\sigma^2/n_i = 2/n_i$.
   c. We generated the variance of the trial outcome based on a χ^2 distribution with $2n_i$-2 degrees of freedom, divided by n_i-1.

3. For each series:

A DL analysis and an HKSJ analysis were carried out.

4. For each scenario, I^2 and each meta-analysis method, we calculated the error rate, i.e. the percentage of series that had a statistically significant ($p<0.05$) outcome.

B. The simulations for the odds ratio

1. When the outcome was dichotomous, we had to choose an additional parameter: the overall event rate p_0. We varied the p_0 between 0.1 and 0.9 and for each value we used (2) to calculate τ^2, with

   $$\epsilon^2 = \frac{1}{k} \sum \frac{1}{n_i} \left(\frac{2}{p_0} + \frac{2}{1-p_0} \right)$$

2. For each trial i:
   a. We determined the 'true' trial effect size $\ln(\text{odds ratio}_i)=\delta_i$, where δ_i was a random draw from the normal distribution with mean 0 and variance τ^2.
   b. We calculated the event rates p_a and p_b in the two groups, such that: $\ln(p_a /(1- p_a)) = \ln(p_0 /(1- p_0)) - \delta_i/2$, and $\ln(p_b /(1-p_b)) = \ln(p_0 /(1- p_0)) + \delta_i/2$.
   c. We generated the observed event rates P_a and P_b in each group based on Bernouilli distributions with event rates p_a and p_b, respectively.
   d. Based on P_a and P_b, we calculated the natural log of the odds ratio and its variance $(1/P_a +1/(1- P_a) +1/P_b +1/(1- P_b))/n_i$.

Steps 3 and 4 were the same as for a continuous outcome.

C. The simulations for the risk ratio

The risk ratio simulation was similar to the odds ratio simulation, but the variance was different:

$$\epsilon^2 = \frac{1}{k}\sum\frac{1}{n_i}\left(\frac{2}{p_0} - 2\right)$$

Furthermore, for each trial:

a. We determined the 'true' trial risk ratio ln(risk ratio$_i$)= δ_i , where δ_i was a random draw from the normal distribution with mean 0 and variance τ^2.

b. We calculated the event rates p_a and p_b in the two groups, such that: $\ln(p_a) = \ln(p_0) - \delta_i/2$ and $\ln(p_b) = \ln(p_0) + \delta_i/2$. Event rates below 0.01 or above 0.99 were replaced by 0.01 or 0.99, respectively.

c. We generated the observed event rates P_a and P_b in each group based on Bernouilli distributions with event rates p_a and p_b, respectively.

d. This led to the natural log of the risk ratio and its variance $(1/P_a + 1/P_b - 2)/n_i$.

Appendix 3: R code for the conversion of DL to HKSJ results

The R package metafor [20] can also be used to perform an HKSJ analysis. The implementation is based on the meta-regression paper by Knapp and Hartung [34]: when no covariates or moderator variables are used, the meta-regression reduces to a random effects meta-analysis as proposed by Hartung/Knapp and Sidik/Jonkman.

The usual approach to perform an HKSJ analysis with metafor is based on study effects combined with fixed effects weights or standard errors. In our examples the HKSJ method must be applied on random effects weights instead of fixed effects weights. This can be done by choosing a fixed effects analysis (method="FE") in combination with the HKSJ method. This will result in warnings, because in general the HKSJ adjustment is not meant to be used in combination with a fixed effects analysis. In this case, the warnings can be neglected. The code is kindly provided by G Knapp.

Code for HKSJ conversion in R: first example

```
library(metafor)

  y <- c(−0.04, -0.07, -0.31, -1.36, -0.54)
  w <- c( 24.0, 22.2, 21.3, 15.5, 17.0)
  rma.uni(y, vi = 1/w, method="FE", knha=TRUE)
```

Output is presented in Table 5.

Table 5 R output for first example (Hartung-Knapp-Sidik-Jonkman method)

	Estimate	SE	t-value	p-value	CI.LB	CI.UB
R output	−0.3938	0.2254	−1.7473	0.1555	−1.0195	0.2319

Relevant part from output from R package metafor. SE: standard error; CI.LB: lower bound of 95% confidence interval; CI.UB: upper bound of 95% confidence interval.

Code for HKSJ conversion in R: second example (ln HR)

```
library (metafor)

  y <- c(0.81, 0.67, 0.80, 0.91, 0.56, 0.98, 1.24, 0.75, 0.95, 0.66)
  w <- c( 5.0, 2.1, 11.5, 46.7, 2.9, 9.3, 3.9, 12.7, 3.9, 2.0)
  # meta-analysis on log scale (ln HR). Note the brackets around the following
  # syntax!
  (hr <- rma.uni(log(y), vi=1/w, method="FE", knha=TRUE) )
  # backtransformation:
  exp(hr$b)
  exp(c(hr$ci.lb, hr$ci.ub)) (Table 6)
```

Table 6 R output for second example (Hartung-Knapp-Sidik-Jonkman method)

	Estimate (HR)	SE	t-value	p-value	CI.LB	CI.UB
R output	−0.1458	0.0470	−3.1031	0.0127	−0.2521	−0.0395
After back-transformation	0.8643				0.7772	0.9613

Relevant part from output from R package metafor. HR: hazard ratio; SE: standard error; CI.LB: lower bound of 95% confidence interval; CI.UB: upper bound of 95% confidence interval.

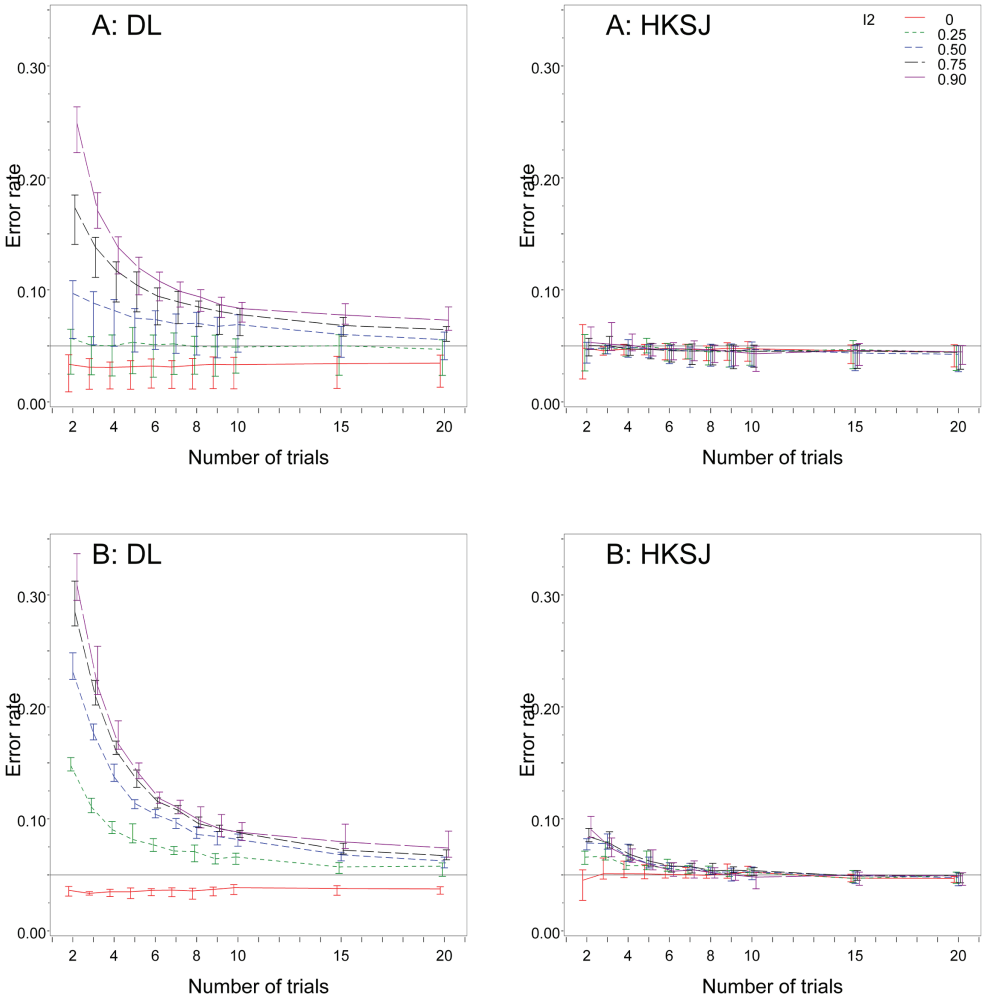

Figure S1 (A and B). DerSimonian-Laird and Hartung-Knapp-Sidik-Jonkman error rates for Risk Ratios, for various I^2 and mixtures of trial sizes: A: Equally sized trials; B: One small trial, 1/10th of other trials. Vertical bars refer to the minimum and maximum error rates over the group sizes. The lines connect the means of these error rates. DL: DerSimonian & Laird meta-analysis method. HKSJ: Hartung-Knapp-Sidik-Jonkman meta-analysis method.

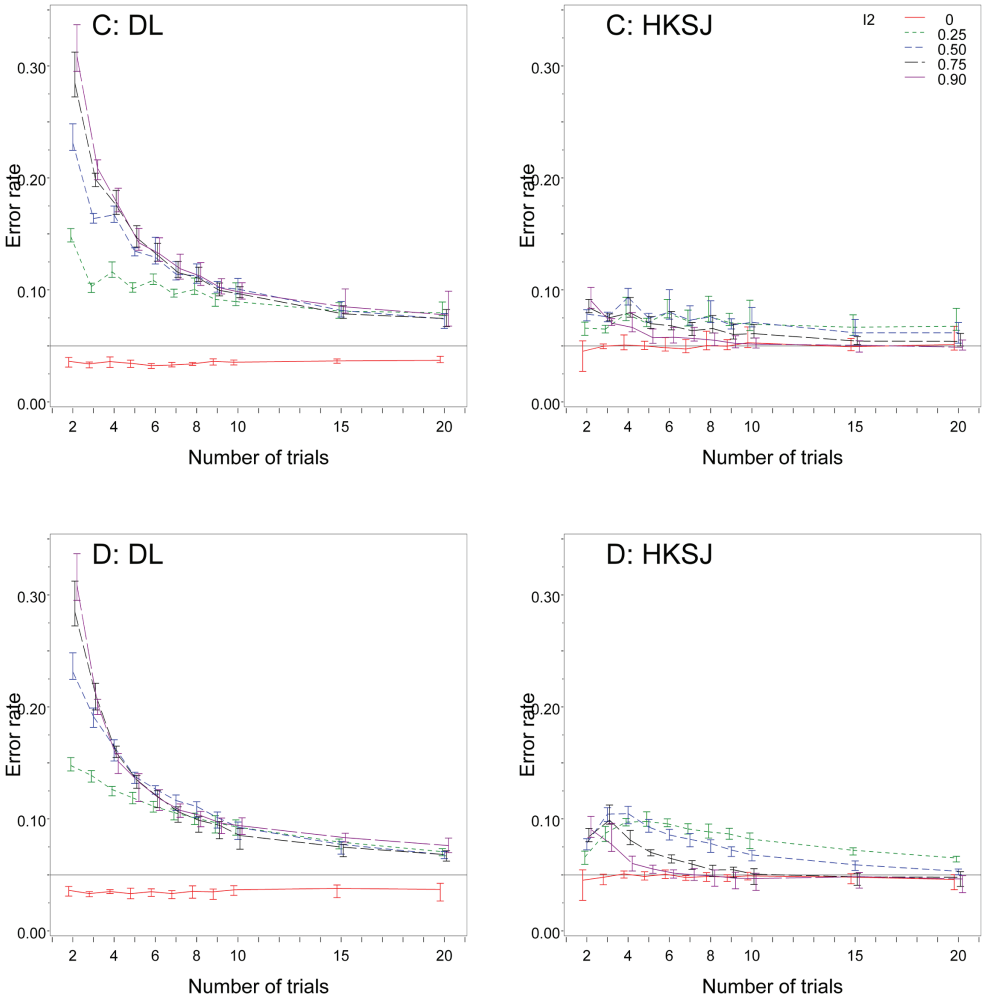

Figure S1 (C and D). DerSimonian-Laird and Hartung-Knapp-Sidik-Jonkman error rates for Risk Ratios, for various I^2 and mixtures of trial sizes: C: 50-50 small and large trials (ratio 1:10); D: one large trial (10 times larger than other trials). Vertical bars refer to the minimum and maximum error rates over the group sizes. The lines connect the means of these error rates. DL: DerSimonian & Laird meta-analysis method. HKSJ: Hartung-Knapp-Sidik-Jonkman meta-analysis method.

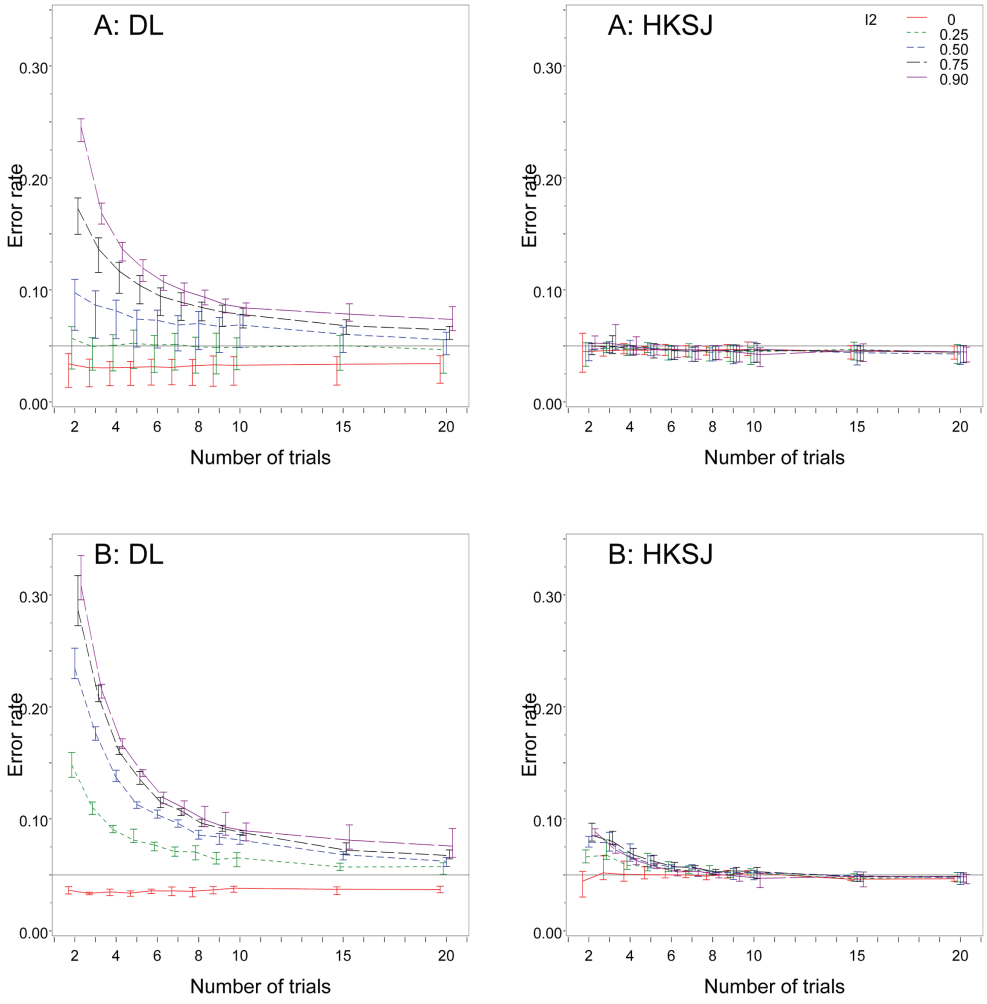

Figure S2 (A and B). DerSimonian-Laird and Hartung-Knapp-Sidik-Jonkman error rates for Odds Ratios, for various I^2 and mixtures of trial sizes: A: Equally sized trials; B: One small trial, 1/10th of other trials. Vertical bars refer to the minimum and maximum error rates over the group sizes. The lines connect the means of these error rates. DL: DerSimonian & Laird meta-analysis method. HKSJ: Hartung-Knapp-Sidik-Jonkman meta-analysis method.

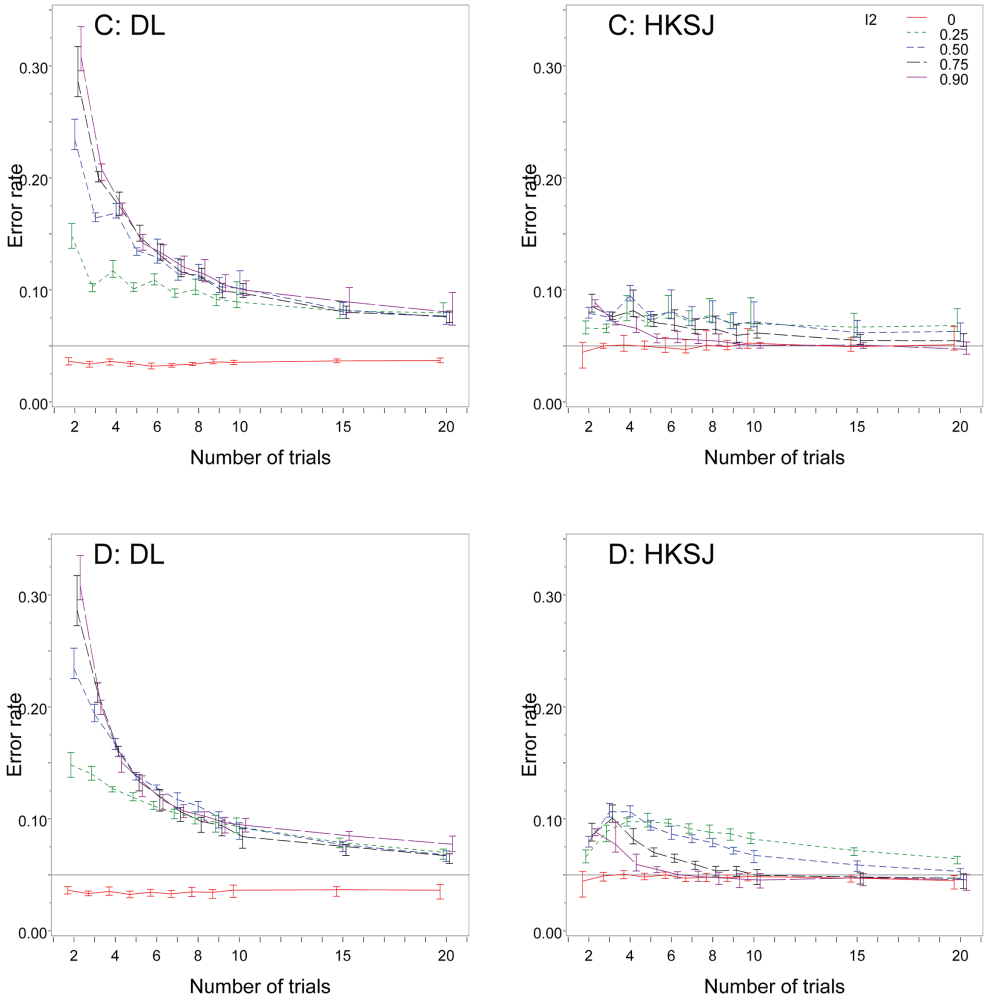

Figure S2 (C and D). DerSimonian-Laird and Hartung-Knapp-Sidik-Jonkman error rates for Odds Ratios, for various I^2 and mixtures of trial sizes: C: 50-50 small and large trials (ratio 1:10); D: one large trial (10 times larger than other trials). Vertical bars refer to the minimum and maximum error rates over the group sizes. The lines connect the means of these error rates. DL: DerSimonian & Laird meta-analysis method. HKSJ: Hartung-Knapp-Sidik-Jonkman meta-analysis method.

Conversion of DL to HKSJ for continuous outcomes

Note: Adapt values in the grey cells

Calculation of t-value	
Confidence level for 95% conf. interval	95
Number of studies	5
Corresponding t-value	2.77644511
Reported DL pooled effect	-0.39
Number of decimals required in Confidence Interval	2

	Reported DL results		**Calculated HKSJ weights**	
study	**weight W_i**	**study effect y_i**	$(y_i - y)^2$	$w_i * (y_i - y)^2$
1	24	-0.04	0.1225	2.94
2	22.2	-0.07	0.1024	2.27328
3	21.3	-0.31	0.0064	0.13632
4	15.5	-1.36	0.9409	14.58395
5	17	-0.54	0.0225	0.3825
6			0	0
7			0	0
8			0	0
9			0	0
10			0	0
11			0	0
12			0	0
13			0	0
14			0	0
15			0	0
Sum	100			20.31605

HKSJ Results

SE HKSJ	0.2253666
Point estimate	-0.39
CI_lower	-1.02
CI_upper	0.24
t-value	-1.7305134
degrees of freedom	4
p-value	0.1585873

Figure S3.A Excel template for the conversion of DL to HKSJ results for continuous outcomes, Example 1. SE: Standard error; DL: DerSimonian-Laird result; HKSJ: Hartung-Knapp-Sidik-Jonkman result; CI: Confidence Interval. Web material available on http://www.biomedcentral.com/1471-2288/14/25/additional.

Conversion of DL to HKSJ for ratio outcomes

Note: Adapt values in the grey cells

Calculation of t-value	
Confidence level	95
Number of studies	10
Corresponding t-value	2.262157158
Reported DL pooled effect	0.86
Pooled DL effect on ln-scale	-0.15082289
Number of decimals required in Conf.Interval	2

	Reported DL results			**Calculated HKSJ weights**	
study	**weight w_i**	**study effect y_i**	**ln(study effect) $\ln(y_i)$**	***$(\ln(y_i)-\ln(y))$***	***w_i^* $(\ln(y_i)-\ln(y))^2$***
1	5	0.81	-0.210721031	0.00358778	0.017938937
2	2.1	0.67	-0.400477567	0.06232745	0.130887661
3	11.5	0.8	-0.223143551	0.00523027	0.060148198
4	46.7	0.91	-0.094310679	0.00319363	0.149142517
5	2.9	0.56	-0.579818495	0.18403723	0.533707966
6	9.3	0.98	-0.020202707	0.01706163	0.158673178
7	3.9	1.24	0.21511138	0.13390788	0.522240769
8	12.7	0.75	-0.287682072	0.01873043	0.237876536
9	3.9	0.95	-0.051293294	0.00990614	0.038633947
10	2	0.66	-0.415515444	0.07006214	0.140124297
11			0	0	0
12			0	0	0
13			0	0	0
14			0	0	0
15			0	0	0
Sum	100				1.989374005

HKSJ Results

SE HKSJ	0.0470151
Point estimate	0.86
CI_lower	0.77
CI_upper	0.96
t-value	-3.20797
degrees of freedom	9
p-value	0.0106937

Figure S3.B Excel template for the conversion of DL to HKSJ results for ratio outcomes, Example 2. SE: Standard error; DL: DerSimonian-Laird result; HKSJ: Hartung-Knapp-Sidik-Jonkman result; CI: Confidence Interval. Web material available on http://www.biomedcentral.com/1471-2288/14/25/additional.

This excel book contains supplemental information to the following paper:
The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method.
Joanna IntHout, John P.A. Ioannidis, George F. Borm.
BMC Medical Research Methodology, 2014

Contact details
Joanna IntHout
Radboud university medical center,
Biostatics,
Department for Health Evidence,
Nijmegen, The Netherlands
Joanna.IntHout@radboudumc.nl

Disclaimer
This excel book was created with Excel 2007. Please contact the author in case of problems.

Figure S3.C Information and disclaimer for Excel template for the conversion of DL to HKSJ results. Web material available on http://www.biomedcentral.com/1471-2288/14/25/additional.

References

1. DerSimonian R, Laird N: Meta-analysis in clinical trials. Control Clin Trials 1986, 7(3):177–188.
2. Collaboration TC: Review Manager (RevMan) 5.1.4. Copenhagen: The Nordic Cochrane Centre; 2011.
3. Borenstein M, Hedges L, Higgins J, Rothstein H: Comprehensive Meta-analysis Version 2. Englewood NJ: Biostat; 2005.
4. Hartung J: An alternative method for meta-analysis. Biom J 1999:901–916.
5. Hartung J, Knapp G: A refined method for the meta analysis of controlled clinical trials with binary outcome. Stat Med 2001, 20(24):3875–3889.
6. Hartung J, Knapp G: On tests of the overall treatment effect in meta analysis with normally distributed responses. Stat Med 2001, 20(12):1771–1782.
7. Follmann DA, Proschan MA: Valid inference in random effects meta-analysis. Biometrics 1999, 55(3):732–737.
8. Hartung J, Makambi KH: Reducing the number of unjustified significant results in meta-analysis. Commun Stat Simul Comput 2003, 32(4):1179–1190.
9. Makambi KH: The effect of the heterogeneity variance estimator on some tests of treatment efficacy. J Biopharm Stat 2004, 14(2):439–449.
10. Sidik K, Jonkman JN: Robust variance estimation for random effects meta-analysis. Comput Stat Data Anal 2006, 50(12):3681–3701.
11. Sidik K, Jonkman JN: A simple confidence interval for meta-analysis. Stat Med 2002, 21(21):3153–3159.
12. Sidik K, Jonkman JN: On constructing confidence intervals for a standardized mean difference in meta-analysis. Commun Stat Simul Comput 2003, 32(4):1191–1203.
13. Sánchez-Meca J, Marín-Martínez F: Confidence intervals for the overall effect size in random-effects meta-analysis. Psychol Meth 2008, 13(1):31.
14. Sidik K, Jonkman JN: Simple heterogeneity variance estimation for meta analysis. J Roy Stat Soc 2005, 54(2):367–384.
15. Davey J, Turner R, Clarke M, Higgins J: Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: a cross-sectional, descriptive analysis. BMC Med Res Methodol 2011, 11(1):160.

16. Higgins J, Thompson SG, Deeks JJ, Altman DG: Measuring inconsistency in meta-analyses. Bmj 2003, 327(7414):557.
17. Ioannidis JP, Patsopoulos NA, Evangelou E: Uncertainty in heterogeneity estimates in meta-analyses. Bmj 2007, 335(7626):914–916.
18. IntHout J, Ioannidis JP, Borm GF: Obtaining evidence by a single well-powered trial or several modestly powered trials. Stat Methods Med Res 2012: [Epub ahead of print].
19. Borm GF, Lemmers O, Fransen J, Donders R: The evidence provided by a single trial is less reliable than its statistical analysis suggests. J Clin Epidemiol 2009, 62(7):711–715. e711.
20. Viechtbauer W: Conducting meta-analyses in R with the metafor package. J Stat Softw 2010, 36(3):1–48.
21. Harbord RM, Higgins JP: Meta-regression in Stata. Meta 2008, 8(4):493–519.
22. Higgins JPT, Thompson SG: Quantifying heterogeneity in a meta-analysis. Stat Med 2002, 21(11):1539–1558.
23. The Cochrane Collaboration. http://www.cochrane.org/cochrane-reviews.
24. Singh M, Das RR: Zinc for the common cold. Cochrane Database Syst Rev 2011, 2:CD001364.
25. Pidala J, Djulbegovic B, Anasetti C, Kharfan-Dabaja M, Kumar A: Allogeneic hematopoietic cell transplantation for acute lymphoblastic leukemia (ALL) in first complete remission. Cochrane Library 2011, 10:CD008818.
26. Thorlund K, Wetterslev J, Awad T, Thabane L, Gluud C: Comparison of statistical inferences from the DerSimonian–Laird and alternative random-effects model meta-analyses – an empirical assessment of 920 Cochrane primary outcome meta-analyses. Res Synth Meth 2011, 2(4):238–253.
27. Kontopantelis E, Reeves D: Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: A simulation study. Stat Methods Med Res 2012, 21(4):409–426.
28. Brockwell SE, Gordon IR: A simple method for inference on an overall effect in meta-analysis. Stat Med 2007, 26(25):4531–4543.
29. Hardy RJ, Thompson SG: A likelihood approach to meta-analysis with random effects. Stat Med 1996, 15(6):619–629.
30. Shuster JJ: Empirical vs natural weighting in random effects meta-analysis. Stat Med 2010, 29(12):1259–1265.
31. Henmi M, Copas JB: Confidence intervals for random effects meta analysis and robustness to publication bias. Stat Med 2010, **29**(29)**:**2969–2983.

32. Guolo A: Higher-order likelihood inference in meta-analysis and meta-regression. Stat Med 2012, **31**(4)**:**313–327.
33. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR: Introduction to Meta-Analysis. Chichester, UK: Wiley; 2009.
34. Knapp G, Hartung J: Improved tests for a random effects meta-regression with a single covariate. Stat Med 2003, **22:**2693–2710.

Chapter 5

Small studies are more heterogeneous than large ones: a meta-meta-analysis

IntHout J, Ioannidis JP, Borm GF, Goeman JJ.
Journal of Clinical Epidemiology. 2015;68(8):860-9.

Abstract

Objective

Between-study heterogeneity plays an important role in random-effects models for meta-analysis. Most clinical trials are small, and small trials are often associated with larger effect sizes. We empirically evaluated whether there is also a relationship between trial size and heterogeneity (τ).

Study Design and Setting

We selected the first meta-analysis per intervention review of the Cochrane Database of Systematic Reviews Issues 2009-2013 with a dichotomous (n=2009) or continuous (n=1254) outcome. The association between estimated τ and trial size was evaluated across meta-analyses using regression and within meta-analyses using a Bayesian approach. Small trials were predefined as those having standard errors over 0.2 standardized effects.

Results

Most meta-analyses were based on few (median 4) trials. Within the same meta-analysis, the small-study τ_S^2 was larger than the large-study τ_L^2 (average ratio 2.11, 95% Credible Interval (1.05, 3.87) for dichotomous and 3.11 (2.00, 4.78) for continuous meta-analyses). The imprecision of τ_S was larger than of τ_L: median SE 0.39 vs. 0.20 for dichotomous and 0.22 vs. 0.13 for continuous small-study and large-study meta-analyses.

Conclusion

Heterogeneity between small studies is larger than between larger studies. The large imprecision with which τ is estimated in a typical small-studies' meta-analysis is another reason for concern and sensitivity analyses are recommended.

What is new?

Key findings

- In a sample of 2009 meta-analyses with a dichotomous outcome and 1254 meta-analyses with a continuous outcome of the CDSR Issues 2009-2013, the between-study heterogeneity τ was often estimated to be either zero or high, and the imprecision of the estimated τ was large, especially for meta-analyses based on few and/or small studies.
- Small studies had higher mean heterogeneity estimates than medium/large studies of the same meta-analysis.

What this adds to what was known?

- Evidence from small studies tends to show not only larger effect sizes but also larger and less precise estimates of between-study heterogeneity.

What is the implication and what should change now?

- In a random-effects meta-analysis the estimated between-study heterogeneity directly affects the summary treatment effect and prediction interval. It should be realized that the estimated τ is often imprecise and on average larger for small studies. Sensitivity analyses to check robustness of the pooled effect estimate may be warranted.

1. Introduction

In clinical research many small and possibly underpowered studies are conducted. Among interventional trials registered between 2007 and 2010 in ClinicalTrials.gov, 62% (17,726/28,458) enrolled at most 100 participants[1]. In 2008, 70% (10,492/14,886) of the meta-analyses with a binary outcome in the Cochrane Database of Systematic Reviews (CDSR), Issue 1, consisted only of studies with less than 50% power to detect a 30% relative risk reduction[2].

There is an ongoing debate on the disadvantages of small trials[3]. Small trials are associated with larger treatment effect estimates[2, 4, 5], and it is possible that between-study heterogeneity also increases when studies are smaller. Turner et al[2] observed that removing the underpowered (<50% power) studies from 1107 meta-analyses resulted in a median 21% decrease in the estimated τ^2. Borm et al[6] observed higher heterogeneity between small rheumatoid arthritis studies compared to larger studies. Individual study results are influenced by many, possibly related aspects, like quality of study, publication bias and study size[7-10]. Califf et al[1] observed that small trials contain significant heterogeneity in methodological approaches, including reported use of randomization, blinding, and data monitoring committees. Button et al[11] argued that underpowered studies are prone to several analytical and reporting biases. Small studies may be of lower quality in other aspects of their design as well. This may affect the between-study heterogeneity.

The current paradigm, in which multiple small studies are conducted and subsequently combined in a meta-analysis, is questioned[3, 12]. Especially in random-effects models and with substantial heterogeneity, the influence of small studies will be major and may affect the reliability of meta-analyses. On the other hand, simulations have shown that a meta-analysis containing many, possibly small, studies, is better than a single large trial able to estimate the treatment effect[6, 13, 14], even when there is some publication bias[15]. Roloff et al. [16] showed that in case of cumulative meta-analysis it is more powerful to add several small studies than one or a few large studies, because the between-study heterogeneity can be estimated more precisely when more studies, either small or large, are available. However, a questionable assumption underlying their calculations is that heterogeneity is similar between small and large studies. The same questionable assumption occurs in standard applications of

random-effects meta-analysis: one single τ^2 is used in the random-effects weights for all studies.

If there is heterogeneity, treatment effects in individual studies may deviate more from the summary effect than expected by chance. Simulations have shown that when there is heterogeneity but no true treatment effect, the frequency of false statistically significant findings in single trials increases more than 10-fold[15]. When small studies have higher than average heterogeneity, the increase in error rates for small single trials will be even larger. Also, prediction intervals[17] constructed with an average τ will result in too narrow predictions for the expected effect for future small trials.

In summary, if there is a difference in heterogeneity between small and large trials, this can influence both the reliability of the results of single trials and of meta-analyses. Results of the current method for random-effects meta-analysis may be overly drawn towards the small-study results, prediction intervals may be too narrow, and false-positive findings of single trials may occur more frequently than expected.

In this paper we investigate empirically whether the heterogeneity of small and large trials is different. We used meta-analyses from 3851 reviews on interventions of the 2009-2013 Issues from the Cochrane Database of Systematic Reviews (CDSR). First we investigated in a cross-sectional approach the relation between study size and heterogeneity across 3263 meta-analyses. As Turner et al. [18] showed that the extent of heterogeneity could be related to outcome and intervention type, our primary analysis is a paired-data approach, comparing the between-study heterogeneity of large trials with the small-study heterogeneity of the same meta-analysis.

2. Methods

2.1 Selected data

The UK Cochrane Editorial Unit provided us with the statistical data of the systematic reviews of interventions, included in the CDSR Issues of 2009-2013. We used the mean values and standard deviations per treatment group for meta-analyses with continuous outcomes, and counts (with/without event) for those with dichotomous outcomes. Most Cochrane reviews included multiple meta-analyses, and meta-analyses from the same review are often correlated. The first reported analysis in a review is usually one of the primary analyses. Hence, to avoid subjectivity in selecting specific meta-analyses we used only the first meta-analysis appearing in the Data and Analyses section that was based on at least two studies. In order to maximize the number of meta-analyses for our evaluation, we used both the first continuous and the first binary outcome meta-analysis, if available. A selected meta-analysis could contain sub-analyses. These sub-analyses were not always mutually exclusive, e.g. when subgroup analyses were done and the same individuals were used in more than one subgroup. If the sub-analyses were combined and resulted in a summary effect-size in the original review we also combined the subgroups. Otherwise, we selected the subgroup analysis based on the largest number of studies.

2.2 Estimation of heterogeneity

Heterogeneity can be expressed as the between-study variance τ^2, with τ in the same unit as the meta-analysis outcome, or as a relative measure I^2, the relative degree of inconsistency across studies [19, 20]. We focused on τ^2, because in the random-effects analysis[21] and also in the reliability of the results of a single trial[13], τ^2 plays a direct role as opposed to I^2, which - even if τ^2 remains the same - is expected to increase with increasing sample sizes of the studies: $E(I^2) = \tau^2/(\tau^2+\sigma^2)$, where σ^2 is the typical within-study variance[19]. Further, the relation between I^2 and sample size already has been investigated by others[22, 23]. We used the empirical Bayes estimator for τ^2, equivalent[24] to the robust[25] Paule and Mandel[25-27] estimator and more accurate than the widely used Method of Moments estimator of DerSimonian and Laird (DL)[21, 28], which is based on large sample assumptions[27]. Further, the relative bias is similar for small to large heterogeneity, as opposed to the DL and Restricted Maximum Likelihood estimators, that underestimate τ^2 more as the heterogeneity increases, especially for dichotomous outcomes[25, 28]. We estimated τ^2 for all

meta-analyses, even when the authors originally performed a fixed-effects analysis. Regression analyses were done with τ, because τ^2 has a very skewed distribution. The paired comparisons between heterogeneity in large- and small-study meta-analyses (section 2.5) were based on Bayesian estimates of τ.

In order to remove unnecessary variation in the outcomes, continuous outcomes were analyzed as standardized mean differences (SMDs). Dichotomous outcomes were analyzed as log Odds Ratios (ORs), where log represents the natural logarithm. If no events were observed in one or both arms, we added 0.5 to all cells of such a study before we estimated the heterogeneity. Adding 0.5 will slightly move the treatment effect to nil and decrease the between-study heterogeneity[29]. As small studies more often have zero events than large studies, this is a conservative approach. Effect sizes were tuned, i.e. in order to give all meta-analyses a positive summary effect (SMD>0 or OR >1), the treatment groups in a meta-analysis were switched if needed. Pooled effect sizes, τ^2 estimates and its standard errors (SEs) resulted from meta-analyses performed with the metafor package version 1.9-2 [30], R software [31] version 3.0.1. The Bayesian approach was carried out using WinBUGs version 1.4.3[32], called from R with the package R2WinBUGS version 2.1-19 [33]. Other statistical analyses were performed with SAS/STAT® software version 9.2 for Windows, copyright© 2002-2008 by SAS Institute Inc., Cary, NC, USA.

2.3 Definition of trial size

We categorized trial size based on trial precision (1/SE of the treatment effect). For this categorization (only) we converted the precision of the trials with an OR outcome into the same order of magnitude as the trials with an SMD by multiplying the precision with $\pi/\sqrt{3}=1.81$[34]. Trials were a priori categorized as very small (size<3), small (3≤size<5), medium sized (5≤size<7.5) or large (size≥7.5). The cutoff size for separating very small/small trials from medium/large trials was set at 5, corresponding to an SE of 0.2 standardized effects.

2.4 Heterogeneity and trial size - across meta-analyses

We categorized the (geometric) mean trial size per meta-analysis in the same way as the trial size (section 2.3). Grouped by mean trial size, the estimated τ was cross-tabulated in three categories (0, 0-0.5, and >0.5), and medians and interquartile ranges (25^{th}-75^{th} percentile, i.e. Q1-Q3) of τ were presented for those meta-analyses with an estimated $\tau>0$. We explored the relation between τ (continuous) and mean log trial precision (1/SE) with weighted linear regression, as the data appeared to be heteroscedastic. We conducted univariable regression and multivariable regression adjusted for the log of the number of studies and Cochrane group, where Cochrane groups with less than seven observations were pooled in an 'other' group. Weights were based on the SE^2 of τ, where the SE(τ) was estimated by taking the squared root of the limits of the 95% confidence interval (CI) for τ^2, defined as $\tau^2 \pm 1.96$ SE, and dividing the distance between these roots by 3.92. In addition, we explored the association between the estimated τ and trial precision per Cochrane group, and also the association with other meta-analysis characteristics (geometric mean study N, ratio of largest/smallest study, effect size, overall precision, total number of subjects and events, and mean event rate), adjusted for the log of the number of studies and Cochrane group.

2.5 Heterogeneity and trial size – paired analyses

An analysis of the association between heterogeneity and trial size across meta-analyses may be hampered by confounders. For example, the endpoint of the analysis may have an impact on both the heterogeneity and the study precision[2, 18]. Hence an apparent relationship between heterogeneity and trial size may be solely the result of differences in endpoints across meta-analyses. A comparison of the heterogeneity of the largest studies versus heterogeneity of the smallest studies of the same meta-analysis is less prone to confounding. Therefore we also investigated the association between trial size (defined as in section 2.3) and heterogeneity within meta-analyses. This has the advantage that the results are matched, i.e. controlled for outcome. In the paired analyses small studies, categorized as having very low or low size (<5), were compared with large studies, i.e. medium or high size (≥5). In sensitivity analyses we restricted the comparison to the very small (<3) versus large trials (≥7.5). We selected meta-analyses that contained at least two trials of each type, i.e. at least two smaller and two larger trials and compared the heterogeneity between the larger studies with the heterogeneity between the smaller studies. If more studies were

available, those were also used in the paired comparison, because heterogeneity estimates are more precise when they are based on larger and/or more studies[35].

In addition to several other characteristics, we estimated per meta-analysis the heterogeneity τ_S for the selected small studies, and in a similar way τ_L of the selected large studies. Effect sizes were tuned, i.e. in order to give the mean of the large-study and small-study meta-analysis a positive summary effect (SMD>0 or OR >1), the treatment groups were switched if needed. Results were summarized with medians and Q1-Q3. In order to compare τ_S and τ_L while incorporating the imprecision of the heterogeneity estimates we used a Bayesian approach. The ratio between the small-study τ_S versus the τ_L of the corresponding larger-study meta-analysis was estimated with Markov Chain Monte Carlo (MCMC) procedures in WinBUGS, using a bivariate lognormal distribution for τ_S and τ_L. The information contained in the posterior distributions for the mean ratio and the τ_S and τ_L was summarized by means and Bayesian (equal-tail) 95% credible intervals (CrI), and indicative p-values based on the percentage of times that the mean ratio of τ_S versus τ_L was ≤1. For the MCMC procedures we took 200,000 iterations with 50,000 for burn-in, and thinning 1 per 100. We checked that the MCMC procedures had reached convergence by visually inspecting the history trace plots, the autocorrelation plots and the cumulative quantile plots for irregularities. The WinBUGS syntax of our model is provided in Table S1 in the appendix.

3 Results

3.1 Selected meta-analyses

In total, 3851 reviews were retrieved from the CDSR Issues of 2009 to 2013 (Figure 1). Selection of reviews containing a meta-analysis with a dichotomous or continuous outcome and based on at least two studies resulted in 2309 reviews: 2009 contain at least one meta-analysis with a dichotomous outcome and 1254 at least one with a continuous outcome that could be used for the across-reviews analyses.

Most meta-analyses were based on a few small studies. Of the 3,263 selected meta-analyses, 1,025 (31%) were based on two studies, 1,226 (38%) on three to

five studies, 603 (18%) on six to ten studies, and 409 (13%) on more than ten studies. The median number of studies per meta-analysis was 4 (Q1-Q3: 2-6), and most of the studies were small: overall, of the 20,185 trials 14,985 (74%) were very small (45%) or small (29%), and only 5,200 (26%) were categorized as larger: 14% as medium sized and 11% as large. Meta-analyses with a dichotomous outcome had a lower average study precision (1/SE) than those with a continuous outcome (median 1.5; Q1-Q3: 1-2.3 versus 4.0; Q1-Q3: 3.1-5.3). The number of very small and small trials was also higher for the meta-analyses with a dichotomous outcome: 79% versus 66% for those with a continuous outcome. The median estimated $\tau 2$ was 0 (Q1-Q3: 0-0.17) for the meta-analyses with a dichotomous outcome and 0.03 (Q1-Q3: 0-0.21) for those with a continuous outcome. More details of the selected meta-analyses are presented in Table 1.

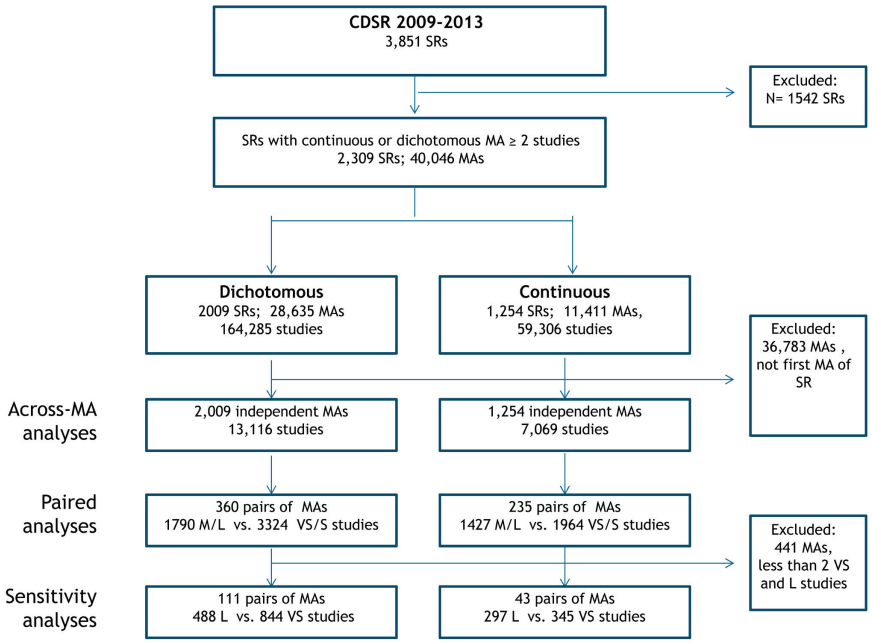

Figure 1. Flowchart of selection and analysis. CDSR 2009-2013: Issues of the Cochrane Database of Systematic Reviews of 2009-2013; SR: systematic review; MA: meta-analysis; VS: very small; S: small; M: medium-sized; L: large.

The total number of Cochrane Groups present in our selection is 52. The Pregnancy and Childbirth Group provided by far most meta-analyses: 229 (11%) with a dichotomous and 135 (11%) with a continuous outcome. An overview of the Cochrane Groups including estimated heterogeneity and study precision can be found in Table S2 in the appendix.

Table 1 Characteristics of the selected meta-analyses

Median (Q1-Q3)	MAs with dichotomous outcome (N=2009)	MAs with continuous outcome (N=1254)	All MAs (N=3263)
MA Effect size (OR / SMD)	1.64 (1.21-2.61)	0.31 (0.14-0.62)	NA
Precision (1/SE) MA effect size	4.3 (2.4-7.7)	7.0 (3.8-13.0)	5.1 (2.7-9.6)
Total number of studies	13,116	7,069	20,185
Total number of small studies [a)]	10,331 (79%)	4,654 (66%)	14,985 (74%)
Number of studies per MA	4 (2-7)	3 (2-6)	4 (2-6)
Number of small studies per MA[a)]	3 (2-6)	2 (1-4)	3 (2-5)
Study precision (1/SE)[b)] per MA	1.5 (1.0-2.3)	4.0 (3.1-5.3)	2.4 (1.3-4.0)
Study N[b)] per MA	99 (58-202)	67 (41-119)	85 (50-164)
Ratio of largest/smallest study per MA	4.3 (2.0-11.3)	3.1 (1.7-7.0)	3.7 (1.9-9.3)
Number of subjects per MA	575 (237-1607)	298 (134-700)	432 (186-1263)
Number of events per MA	114 (40-343)	NA	NA
Mean event rate per MA	0.21 (0.08-0.44)	NA	NA
Estimated τ	0 (0-0.41)	0.16 (0-0.46)	0.07 (0-0.43)
Estimated τ^2	0 (0-0.17)	0.03 (0-0.21)	0 (0-0.18)
I^2 (%)	0 (0-43.1)	39.3 (0-77.3)	4.8 (0-58.9)

Data are summarized with median and interquartile range (Q1-Q3).
MA: meta-analysis; OR: Odds Ratio; SMD: standardized mean difference; SE: standard error; NA: Not Applicable.
[a)] Small studies contain the very small studies (size<3) and the small studies (size<5), see section 2.3 for the definition of size. [b)] geometric mean.

3.2 Heterogeneity versus study size – across meta-analyses

In 1559 (48%) of the 3263 selected meta-analyses the estimated heterogeneity was zero. This percentage was higher for the meta-analyses with a dichotomous than with a continuous outcome: 55% vs. 36%.

Figure 2 shows the estimated τ in relation to the mean study precision. The percentages in Table 2 suggest that in small studies the estimated τ is more often either equal to zero or larger than 0.5. For larger studies, τ seems more of moderate size, i.e. between zero and 0.5. For example, the percentage of dichotomous outcome meta-analyses with a moderate τ increased from 13% for meta-analyses with on average very small studies to 40% for meta-analyses of large studies. A similar increase, from 24% to 59%, is seen for the meta-analyses with a continuous outcome. The number of zero and large τ estimates decreased correspondingly. This pattern is also observed in the meta-analyses with an estimated $\tau > 0$.

The estimated τ was negatively associated with the mean study precision according to the univariable regression (Table 2), which suggests that small studies on average may have larger between-study heterogeneity. For the meta-analyses with a dichotomous outcome, the negative association was no longer significant after adjustment for log number of trials and Cochrane Group, whereas for those with a continuous outcome it was.

Most meta-analysis characteristics were much stronger associated with the estimated τ than the study precision. The increase in model R^2 due to study precision was only 1% and 3% for the meta-analyses with a dichotomous and continuous outcome, respectively, compared to a model with only the Cochrane Group and log number of studies. Not surprisingly, the precision of the meta-analysis effect estimate was by far the most associated, with an increase in model R^2 of 39% and 43%, respectively. The meta-analysis effect size showed the second largest increase in R^2 (10% and 14%, respectively), see Table S3 in the Appendix.

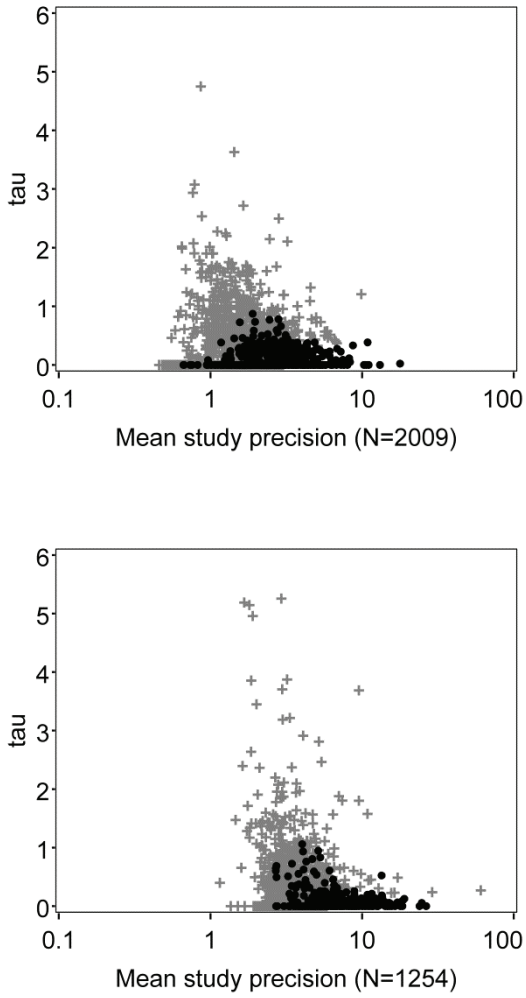

Figure 2. Estimated heterogeneity (τ) versus trial precision across meta-analyses. Left: meta-analyses with a dichotomous outcome; right: meta-analyses with a continuous outcome. Mean study precision is the geometric mean of the precisions (1/SE) of the studies in a meta-analysis. Black dots correspond to meta-analyses with the 25% largest weights in the weighted regression, i.e. the 25% τ estimated most precisely.

Table 2 Estimated τ in relation to study precision across meta-analyses

	Mean study size per MA [a)]			
	Very small	Small	Medium	Large
MAs with dichotomous outcome (N=2009)				
Overall	1108	580	214	107
Estimated τ (n (%))				
τ = 0	731 (66)	255 (44)	73 (34)	50 (47)
0 < τ ≤0.5	144 (13)	196 (34)	109 (51)	43 (40)
τ > 0.5	233 (21)	129 (22)	32 (15)	14 (13)
Estimated τ if >0				
N (MAs with est. τ >0)	377	325	141	57
Median τ	0.64	0.43	0.31	0.27
Q1-Q3	0.37-1.03	0.29-0.65	0.19-0.46	0.15-0.46
P90	1.58	0.96	0.69	0.73
Linear regression[b)]				
Unadjusted slope	-0.041, 95% CI (-0.055, -0.028)			p< 0.001
Adjusted slope	-0.013, 95% CI (-0.028, 0.001)			p= 0.078
MAs with continuous outcome (N=1254)				
Overall	282	603	246	123
Estimated τ (n (%))				
τ = 0	112 (40)	209 (35)	86 (35)	43 (35)
0 < τ ≤0.5	68 (24)	245 (41)	127 (52)	73 (59)
τ > 0.5	102 (36)	149 (25)	33 (13)	7 (6)
Estimated τ if >0				
N (MAs with est. τ >0)	170	394	160	80
Median τ	0.65	0.38	0.21	0.12
Q1-Q3	0.36-1.23	0.23-0.65	0.12-0.41	0.08-0.24
P90	1.81	1.01	0.78	0.47
Linear regression[b)]				
Unadjusted slope	-0.096, 95% CI (-0.113, -0.078)			p< 0.001
Adjusted slope	-0.072, 95% CI (-0.091, -0.052)			p< 0.001

MA: meta-analysis; CI: confidence interval; Q1-Q3: 25th and 75th percentile of distribution; P90: 90th percentile of distribution.

[a)] Based on geometric mean study size per MA (section 2.4). [b)] Results of weighted linear regression with estimated τ as dependent and mean log study precision as independent variable, unadjusted, and adjusted for log number of studies and Cochrane group.

3.3 Trial size versus heterogeneity – paired analyses

The 2009 series of trials with a dichotomous outcome resulted in 360 pairs of at least two large and two small trials; the 1254 series with a continuous outcome in 235 pairs. Table 3.A shows the characteristics of the pairs. We found somewhat larger effect sizes for the meta-analyses based on small trials than on large trials, both for the dichotomous and the continuous outcomes. The imprecision (SE) of the τ estimates was higher for the small-study meta-analyses compared with the large-study meta-analyses, and much higher for the meta-analyses based on very small studies. The number of meta-analyses where τ was estimated to be zero was high, especially in the very small/small-study meta-analyses with a dichotomous outcome: 53%, compared to 42% for the corresponding pairs of medium/large-study meta-analyses, and 30% vs. 29% for the continuous outcome meta-analyses. The occurrence of τ>0.5 estimates also differed between the (very) small and the medium/large-study meta-analyses: 25% versus 11% for the meta-analyses with a dichotomous and 31% versus 9% for those with a continuous outcome. Overall, the estimated τ_S was larger than τ_L, with point estimates of 0.22 vs. 0.15 and 0.27 vs. 0.15, and estimated mean ratios for τ^2_S / τ^2_L of 2.11 (95% CrI 1.05–3.87) for the meta-analyses with a dichotomous and 3.11 (95% CrI 2.00–4.78) for those with a continuous outcome.

We also restricted the comparisons to the more extreme study size differences, see Table 3.B. For the meta-analyses with a dichotomous outcome, the percentage of τ=0 estimates was 67% for the very small studies versus 41% for the large studies, compared to twice 26% for the very small- and large-study meta-analyses with a continuous outcome. The occurrence of τ>0.5 was 26% versus 6% for the very small- and large-study dichotomous and 58% versus 12% for the continuous outcome meta-analyses. For the meta-analyses with a continuous outcome the mean ratio τ^2_S/τ^2_L was again confidently larger than 1 (7.28, 95% CrI 2.76–18.03), but for those with a dichotomous outcome the point estimate was smaller than 1 and the CrI contained the 1 (0.03, 95% CrI 0.00–2.23). Posterior distributions of the mean heterogeneity (τ_S and τ_L) are presented in Figure 3.

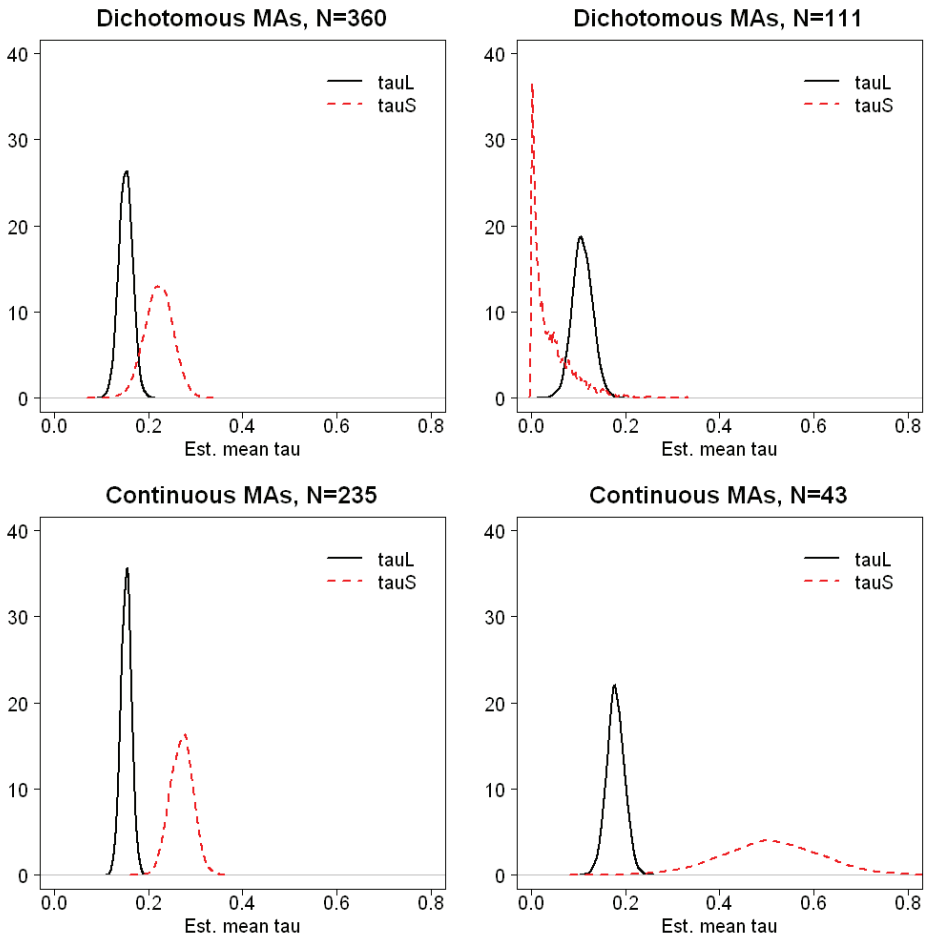

Figure 3. Posterior distributions of the mean heterogeneity between small (τ_S) and large (τ_L) trials resulting from paired comparisons. Posterior densities are obtained with WinBUGS. MAs=meta-analyses. Left column: all possible pairs (very small/small vs. medium/large studies). Right column: pairs of very small vs. large studies.

Table 3.A Heterogeneity versus trial size – paired comparisons – All meta-analyses

Median (Q1-Q3)	Trial size	Dichotomous Very Small/ Small	Dichotomous Medium/ Large	Continuous Very Small/ Small	Continuous Medium/ Large
N		360	360	235	235
No of studies		6 (3-11)	3 (2-5)	5 (3-9)	3 (2-7)
Precision (1/SE) [a]		1.5 (1.2-1.8)	3.9 (3.5-4.7)	3.5 (3.1-3.9)	6.9 (6.0-7.9)
Study N [b]		93 (67-146)	397 (262-734)	52 (41-66)	195 (151-261)
No of subjects		694 (342-16301)	1646 (839-4190)	284 (164-518)	799 (390-2184)
Effect size [c] **(OR/SMD)**		1.72 (1.24-2.53)	1.29 (1.10-1.71)	0.40 (0.20-0.68)	0.20 (0.10-0.38)
No of events		125 (62-241)	412 (241-956)		
Mean event rate		0.20 (0.08-0.42)	0.29 (0.13-0.46)		
τ=0 [d] (n (%))		190 (53%)	152 (42%)	71 (30%)	69 (29%)
τ>0.5 [d] (n (%))		90 (25%)	40 (11%)	73 (31%)	20 (9%)
SE(τ) [d]		0.39 (0.31-0.52)	0.20 (0.15-0.28)	0.22 (0.18-0.35)	0.13 (0.09-0.20)
Comparison of τ_S with τ_L [e]					
N		360		235	
τ (mean (95% CrI))		0.22 (0.16,0.28)	0.15 (0.12, 0.18)	0.27 (0.22, 0.32)	0.15 (0.13, 0.17)
τ_S / τ_L (mean (95% CrI))		1.45 (1.02, 1.97), p=0.02		1.76 (1.41, 2.19), p<0.001	
τ^2_S /τ^2_L (mean (95% CrI))		2.11 (1.05, 3.87), p=0.02		3.11 (2.00, 4.78), p<0.001	

Data are summarized with median and interquartile range (Q1-Q3), unless otherwise mentioned. OR: Odds Ratio; SMD: standardized mean difference; SE: standard error; Q1-Q3: Interquartile range, with 25th and 75th percentile of distribution; CrI: Credible Interval; p: p-values based on CrI, for indicative purposes.

[a] Precision: geometric mean precision (1/SE) of selected large or small studies in the MA. [b] geometric mean. [c] Effect sizes were tuned such that the mean effect size per MA was > 0 for SMDs and > 1 for ORs. [d] Based on Empirical Bayes estimates for τ. [e] Based on Markov Chain Monte Carlo estimates.

Table 3.B Heterogeneity versus trial size – paired comparisons – Very small-study versus large-study based meta-analyses

		Dichotomous		Continuous	
Median (Q1-Q3)	**Trial size**	**Very Small**	**Large**	**Very Small**	**Large**
N		111	111	43	43
No of studies		4 (2-9)	3 (2-5)	3 (2-6)	5 (2-7)
Precision (1/SE) [a]		1 (0.8-1.2)	5.6 (5.0-7.1)	2.5 (2.2-2.6)	9.8 (8.8-10.7)
Study N [b]		79 (50-146)	985 (570-2043)	28 (23-33)	403 (323-512)
No of subjects		431 (216-1233)	3519 (1901-8407)	107 (59-332)	1921 (1294-4007)
Effect size [c] **(OR/SMD)**		1.98 (1.39-3.59)	1.24 (1.05-1.49)	0.95 (0.46-1.43)	0.31 (0.13-0.42)
No of events		42 (24-102)	763 (465-1793)		
Mean event rate		0.11 (0.04-0.32)	0.20 (0.11-0.41)		
τ=0 [d] (n (%))		74 (67%)	45 (41%)	11 (26%)	11 (26%)
τ>0.5 [d] (n (%))		29 (26%)	7 (6%)	25 (58%)	5 (12%)
SE(τ) [d]		0.61 (0.52-0.79)	0.14 (0.11-0.20)	0.46 (0.30-0.83)	0.11 (0.07-0.19)

Comparison of τ_S with τ_L [e]				
N	111		43	
τ (mean (95% CrI))	0.02 (0, 0.15)	0.11 (0.07, 0.15)	0.50 (0.31, 0.71)	0.18 (0.14, 0.22)
τ_S / τ_L (mean (95% CrI))	0.17 (0.00, 1.50), p=0.91		2.80 (1.66, 4.25), p<0.001	
τ^2_S / τ^2_L (mean (95% CrI))	0.03 (0.00, 2.23), p=0.91		7.82 (2.76, 18.03), p<0.001	

Data are summarized with median and interquartile range (Q1-Q3), unless otherwise mentioned. OR: Odds Ratio; SMD: standardized mean difference; SE: standard error; Q1-Q3: Interquartile range, with 25th and 75th percentile of distribution; CrI: Credible Interval; p: p-values based on CrI, for indicative purposes.

[a] Precision: geometric mean precision (1/SE) of selected large or small studies in the MA. [b] geometric mean. [c] Effect sizes were tuned such that the mean effect size per MA was > 0 for SMDs and > 1 for ORs. [d] Based on Empirical Bayes estimates for τ. [e] Based on Markov Chain Monte Carlo estimates.

4 Discussion

In a random-effects meta-analysis both the treatment effects of the individual studies and the between-study heterogeneity play an important role. Small trials are associated with larger effect sizes than large trials. In absence of publication bias, sample size in itself does not bias the outcome of a study [36]. However, there are several reasons why small trials may also have a different level of heterogeneity than larger trials. Only after promising results of the initial, small exploratory trials, larger trials tend to be conducted, possibly in different patient populations[4, 12, 37]. On the other hand, small trials may suffer from lower quality standards and show a diminished effect[38]. These opposite patterns may cause increased heterogeneity[20] which can only be studied with empirical data.

Using over 3000 first meta-analyses from the CDSR Issues 2009-2013 we have investigated whether there is a relationship between study size and heterogeneity. Most meta-analyses were based on few (median 4; Q1-Q3: 2-6) studies. On average, the estimated τ weakly decreased when precision was larger. In order to minimize effects of outcome and intervention type [18], we performed paired analyses, comparing heterogeneity from the smaller with the larger studies originating from the same meta-analysis. The findings show that the meta-analyses of very small/small studies result in significantly higher mean heterogeneity estimates than medium/large studies. Point estimates indicated an estimated mean τ_S^2 that was at least twice as large as τ_L^2.

In 48% of the meta-analyses the estimated τ was zero, in a similar range as previously reported percentages of dichotomous meta-analyses: 49%, based on meta-analyses of at least two trials[20] and 37%, based on at least four studies[39]. Meta-analyses with low precision for the summary effect, i.e. based on few and/or small studies, resulted more frequently in zero or high (>0.5) point estimates for τ, and the corresponding imprecision (SE) of the estimated τ was large. The high occurrence of zero estimates may be caused by the fact that many of the selected meta-analyses were based on a few small studies and that sample variances for small studies are large. When studies have equal sample variances s^2, the τ^2 estimator corresponds to the sum of $(y_i - y)^2/(k-1)$ minus s^2 [21, 25], or zero if the result is negative, where k is the number of studies. Consequently, for imprecise studies with large s^2 the estimate of τ^2 will often be

zero, even when the true heterogeneity is substantial. The sample variances of the small studies with a dichotomous outcome were larger than those of the small studies with a continuous outcome. This may have caused the abundance of zero τ estimates in the meta-analyses with a dichotomous outcome. This may also be the reason that the estimated mean ratio τ^2_S/τ^2_L was below 1 in the sensitivity analyses with dichotomous outcomes, whereas the other ratios were larger than 1. Also for the primary paired comparisons of the meta-analyses with a dichotomous outcome the ratios were lower compared to the ratios of the meta-analyses with a continuous outcome.

Our study has some limitations, of which probably the most important is that we had to evaluate the relation between τ and study size using the estimates of τ instead of the true values. Especially if studies are small, heterogeneity may be present but difficult to estimate, which complicated in particular the analyses with the dichotomous outcomes. In the current setting of medical interventions, most of the estimates were imprecise. This was even more the case in the paired analyses, where we have split the original meta-analyses in a large- and a small-study part. For a stable estimate of I^2 approximately 500 events and 15 studies are needed[40], and the same may be true for τ. But if we had restricted our analyses to meta-analyses with 15 or more studies, the randomness of the sample would have become questionable. Performing the paired comparisons decreased our sample significantly. Therefore we also evaluated the relation between trial precision and heterogeneity across meta-analyses: this gave us the opportunity to include all meta-analyses. A second limitation is that the set of interventions for which a systematic review is conducted is not random. Some reviews may have been conducted for the very reason of conflicting results and between-study heterogeneity, while some others may have avoided meta-analyses specifically because of high heterogeneity[41]. Further, we used both the first dichotomous and the first continuous outcome meta-analysis from a review if available. Our conclusions with respect to the heterogeneity in meta-analyses with continuous and dichotomous outcomes are thus not independent.

We conclude that the between-study heterogeneity for small studies is larger than for large studies, but often the estimate of τ is imprecise and either zero or high (>0.5). In future research the behavior of the estimated heterogeneity in the context of small studies with a dichotomous outcome should be investigated further. When the estimate of τ^2 is imprecise, we recommend sensitivity analyses

to evaluate the robustness of the estimated combined meta-analysis effect, using various τ^2 values. Using models that do not assume a common random effects distribution across studies –e.g. models with different τ^2 estimates in the weights for small and large studies– is another possibility, but this would be difficult to perform in most meta-analyses, given the limited number of studies. Regardless, the results of random-effects meta-analysis need to be interpreted with extra caution, especially when small studies are involved.

References

1. Califf RM, Zarin DA, Kramer JM, Sherman RE, Aberle LH, Tasneem A. Characteristics of clinical trials registered in clinicaltrials.gov, 2007-2010. JAMA. 2012;307(17):1838-47.
2. Turner RM, Bird SM, Higgins JP. The impact of study size on meta-analyses: examination of underpowered studies in Cochrane reviews. PLoS ONE. 2013;8(3):e59202.
3. Hennekens CH, DeMets D. The need for large-scale randomized evidence without undue emphasis on small trials, meta-analyses, or subgroup analyses. JAMA. 2009;302(21):2361-2.
4. Pereira TV, Horwitz RI, Ioannidis JPA. Empirical evaluation of very large treatment effects of medical interventions. JAMA. 2012;308(16):1676-84.
5. Ioannidis JPA. Why most discovered true associations are inflated. Epidemiology. 2008;19(5):640-8.
6. Borm GF, Donders R. A treatment should be evaluated by small trials. Response to letter. J Clin Epidemiol. 2009;62(8):887-9.
7. Wood L, Egger M, Gluud LL, Schulz KF, Jüni P, Altman DG, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ. 2008;336(7644):601.
8. Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. J Clin Epidemiol. 2000;53(11):1119-29.
9. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997;315(7109):629-34.
10. Nüesch E, Trelle S, Reichenbach S, Rutjes AWS, Tschannen B, Altman DG, et al. Small study effects in meta-analyses of osteoarthritis trials: meta-epidemiological study. BMJ. 2010;341.
11. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience. 2013;14(5):365-76.
12. Ioannidis JP, Pereira TV, Horwitz RI. Emergence of large treatment effects from small trials—reply. JAMA. 2013;309(8):768-9.
13. Borm GF, Lemmers O, Fransen J, Donders R. The evidence provided by a single trial is less reliable than its statistical analysis suggests. J Clin Epidemiol. 2009;62(7):711-5 e1.

14. Borm GF, den Heijer M, Zielhuis GA. Publication bias was not a good reason to discourage trials with low power. J Clin Epidemiol. 2009;62(1):47-53.e3.
15. IntHout J, Ioannidis JP, Borm GF. Obtaining evidence by a single well-powered trial or several modestly powered trials. Statistical methods in medical research. 2012(Epub ahead of print).
16. Roloff V, Higgins J, Sutton AJ. Planning future studies based on the conditional power of a meta-analysis. Statistics in Medicine. 2013;32(1):11-24.
17. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to Meta-Analysis. Chichester, UK: Wiley; 2009.
18. Turner RM, Davey J, Clarke MJ, Thompson SG, Higgins JP. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews. International Journal of Epidemiology. 2012;41(3):818-27.
19. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Statistics in Medicine. 2002;21(11):1539-58.
20. Higgins J, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003;327(7414):557.
21. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7(3):177-88.
22. Rucker G, Schwarzer G, Carpenter J, Schumacher M. Undue reliance on I^2 in assessing heterogeneity may mislead. BMC Medical Research Methodology. 2008;8(1):79.
23. Friedrich JO, Adhikari NK, Beyene J. The ratio of means method as an alternative to mean differences for analyzing continuous outcome variables in meta-analysis: a simulation study. BMC Medical Research Methodology. 2008;8(1):32.
24. Viechtbauer W, López-López JA, Sánchez-Meca J, Marín-Martínez F. A Comparison of Procedures to Test for Moderators in Mixed-Effects Meta-Regression Models. Psychological Methods. 2014 (Advance online publication).
25. Novianti PW, Roes KC, van der Tweel I. Estimation of between-trial variance in sequential meta-analyses: A simulation study. Contemporary Clinical Trials. 2014;37(1):129-38.
26. Paule RC, Mandel J. Consensus values and weighting factors. Journal of Research of the National Bureau of Standards. 1982;87(5):377-85.

27. DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: an update. Contemporary Clinical Trials. 2007;28(2):105-14.
28. Sidik K, Jonkman JN. A comparison of heterogeneity variance estimators in combining results of studies. Statistics in Medicine. 2007;26(9):1964-81.
29. Friedrich JO, Adhikari NK, Beyene J. Inclusion of zero total event trials in meta-analyses maintains analytic consistency and incorporates all available data. BMC Medical Research Methodology. 2007;7(1):5.
30. Viechtbauer W. Conducting meta-analyses in R with the metafor package. Journal of Statistical Software. 2010;36(3):1-48.
31. R Core Team. R: A language and environment for statistical computing. Retrieved from http://www.R-project.org/. Vienna, Austria: R Foundation for Statistical Computing; 2014.
32. Lunn D, Thomas A, Best N, Spiegelhalter D. WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. Statistics and Computing. 2000;10(4):325-37.
33. Sturtz S, Ligges U, Gelman AE. R2WinBUGS: a package for running WinBUGS from R. Journal of Statistical Software. 2005;12(3):1-16.
34. Chinn S. A simple method for converting an odds ratio to effect size for use in meta-analysis. Statistics in Medicine. 2000;19(22):3127-31.
35. Huedo-Medina TB, Sánchez-Meca J, Marín-Martínez F, Botella J. Assessing heterogeneity in meta-analysis: *Q* statistic or I^2 index? Psychological methods. 2006;11(2):193.
36. Batterham AM, Hopkins WG. Emergence of large treatment effects from small trials. JAMA. 2013;309(8):768-9.
37. Krum H, Tonkin A. Why do phase III trials of promising heart failure drugs often fail? The contribution of "regression to the truth". Journal of Cardiac Failure. 2003;9(5):364-7.
38. Villar J, Carroli G, Belizán JM. Predictive ability of meta-analyses of randomised controlled trials. The Lancet. 1995;345(8952):772-6.
39. Ioannidis JPA, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analyses. BMJ. 2007;335(7626):914-6.
40. Thorlund K, Imberger G, Johnston BC, Walsh M, Awad T, Thabane L, et al. Evolution of heterogeneity (I^2) estimates and their 95% confidence intervals in large meta-analyses. PLoS ONE. 2012;7(7):e39471.
41. Ioannidis JP, Patsopoulos NA, Rothstein HR. Reasons or excuses for avoiding meta-analysis in forest plots. BMJ. 2008;336(7658):1413.

Appendix

Table S1 Winbugs model for paired comparisons

```
model {
for (i in 1:MAmax)
        {
      logtaui[i,1:2]     ~ dmnorm(logtau[], Rtau[,])
      tauS[i]   <- exp(logtaui[i,1])
      tauL[i]            <- exp(logtaui[j,2])
      precS.tau2 [i]     <- pow(tauS[i],-2)
      precL.tau2 [i]     <- pow(tauL[i],-2)
      mui[i, 1:2]        ~ dmnorm(mu[], R[ , ])
      muS[i]    <- mui[i,1]
      muL[i]    <- mui[i,2]
      for(j in offsetS[i]:(offsetS[i+1]-1))       # select studies of small-study MA i
            { thetaS[j]  ~  dnorm(muS[i], precS.tau2[i])
                  wS[j]      <- 1/varyS[j]
                  yS[j]       ~  dnorm(thetaS[j], wS[j])
            }
      for(j in offsetL[i]:(offsetL[i+1]-1))       # select studies of large-study MA i
            { thetaL[j]  ~  dnorm(muL[i], precL.tau2[i])
                  wL[j]     <- 1/varyL[j]
                  yL[j]     ~  dnorm(thetaL[j], wL[j])
            }
      logdelta[i] <- logtaui[i,1]-logtaui[i,2]  }
# prior distributions
 logtau[1:2]      ~ dmnorm(logtau0[],prec0[,])
 Rtau[1:2,1:2]   ~  dwish(Omega[,],2)
 mu[1:2]          ~  dmnorm(mu0[],prec0[,])
 R[1:2,1:2]       ~  dwish(Omega[,],2)
  }
# mu0: vector(0,0);  logtau0: vector (-1,-1)
# Omega: Identity matrix I 2x2; prec0:  I x 1.0E-6
```

Table S2 Estimated τ per Cochrane Group in the selected meta-analyses across meta-analyses

	Meta-analyses with dichotomous outcome (N=2009)					Meta-analyses with continuous outcome (N=1254)				
Cochrane Group	n	%	Study precision median (Q1-Q3)	Study N median (Q1-Q3)	Estimated τ median (Q1-Q3)	n	%	Study precision median (Q1-Q3)	Study N median (Q1-Q3)	Estimated τ median (Q1-Q3)
Acute Respiratory Infections	84	4.2	1.7 (1.1-2.7)	154 (85-297)	0.12 (0-0.38)	39	3.1	4.3 (3.4-6.3)	82 (48-162)	0.12 (0-0.42)
Airways	94	4.7	1.4 (0.9-2.1)	109 (52-328)	0 (0-0.20)	86	6.9	3.8 (2.7-5.9)	60 (33-142)	0.04 (0-0.27)
Anaesthesia	53	2.6	1.1 (0.8-1.8)	74 (54-121)	0 (0-0.34)	43	3.4	3.9 (3.2-4.8)	68 (43-97)	0.3 (0-0.58)
Back	11	0.5	1.9 (1.8-2.2)	81 (70-249)	0.14 (0-0.70)	19	1.5	4.6 (3.3-5.8)	86 (43-140)	0.11 (0-0.4)
Bone, Joint and Muscle Trauma	44	2.2	1.0 (0.7-1.4)	60 (43-99)	0 (0-0.29)	24	1.9	3.6 (2.6-4.1)	50 (30-67)	0.12 (0-0.67)
Breast Cancer	19	0.9	2.4 (1.7-3.4)	232(128-426)	0.12 (0-0.44)	4	0.3	4.3 (3.7-4.7)	77 (67-95)	0.51 (0.2-1.92)
Childhood Cancer	6	0.3	0.9 (0.7-1.6)	92 (51-287)	0 (0-0.35)	1	0.1	2.6 (2.6-2.6)	26	0.43
Colorectal Cancer	55	2.7	1.4 (0.9-1.9)	92 (63-195)	0 (0-0.46)	29	2.3	4.0 (3.4-4.9)	69 (53-102)	0.53 (0.15-0.93)
Consumers and Communication	13	0.6	2.8 (1.9-3.6)	180 (93-329)	0.29 (0-0.45)	15	1.2	4.7 (3.6-6.2)	86 (53-162)	0.21 (0-0.59)
Cystic Fibrosis and Genetic Disorders	23	1.1	0.8 (0.6-1.5)	61 (49-89)	0 (0-0.19)	29	2.3	3.2 (2.3-3.9)	48 (28-61)	0.34 (0-0.74)
Dementia and Cognitive Improvement	21	1.0	1.7 (1.2-2.4)	93 (63-164)	0 (0-0.42)	27	2.2	4.3 (3.5-5.8)	77 (49-133)	0.06 (0-0.38)
Depression, Anxiety and Neurosis	48	2.4	2.0 (1.5-2.8)	93 (48-157)	0.18 (0-0.50)	46	3.7	3.8 (2.8-5.5)	60 (37-127)	0.16 (0-0.39)
Developmental, Psychosocial and Learning Problems	24	1.2	2.0 (1.4-2.8)	134 (57-245)	0.1 (0-0.51)	37	3.0	3.8 (3.2-5.7)	67 (43-142)	0.16 (0-0.36)
Drugs and Alcohol	25	1.2	1.5 (1.4-1.8)	70 (50-86)	0.19 (0-0.46)	20	1.6	4.1 (3.3-4.7)	68 (47-96)	0.06 (0-0.43)
Ear, Nose and Throat Disorders	26	1.3	1.5 (1.0-1.8)	66 (50-100)	0.2 (0-0.70)	14	1.1	3.7 (3.1-3.9)	56 (40-69)	0.39 (0.23-0.59)
Effective Practice and Organisation of Care	13	0.6	2.5 (1.8-3.4)	200(129-247)	0 (0-0.43)	13	1.0	5.3 (4.7-5.9)	117 (98-139)	0.16 (0.13-0.33)

Table S2 (cont.) Estimated τ per Cochrane Group in the selected meta-analyses across meta-analyses

	Meta-analyses with dichotomous outcome (N=2009)					Meta-analyses with continuous outcome (N=1254)				
Cochrane Group	n	%	Study precision median (Q1-Q3)	Study N median (Q1-Q3)	Estimated τ median (Q1-Q3)	n	%	Study precision median (Q1-Q3)	Study N median (Q1-Q3)	Estimated τ median (Q1-Q3)
Epilepsy	19	0.9	2.3 (1.3-2.7)	138 (65-223)	0 (0-0.35)	1	0.1	4.3 (4.3-4.3)	73	0.28
Eyes and Vision	23	1.1	1.7 (1.3-3)	101 (54-242)	0.04 (0-0.43)	14	1.1	3.9 (3.6-4.4)	61 (51-76)	0 (0-0.11)
Fertility Regulation	27	1.3	1.4 (0.8-2.6)	338 (99-599)	0 (0-0.24)	11	0.9	5.0 (3.6-7.1)	117 (64-211)	0.09 (0-0.69)
Gynaecological Cancer	58	2.9	1.7 (1.1-2.2)	130 (86-182)	0 (0-0.41)	20	1.6	4.0 (3.2-4.9)	70 (43-99)	0.48 (0-0.86)
HIV/AIDS	30	1.5	2.6 (1.4-4.1)	277(116-480)	0.25 (0-0.56)	13	1.0	5.1 (3.0-6.3)	114 (44-160)	0.24 (0.12-0.77)
Haematological Malignancies	23	1.1	1.9 (1.3-2.6)	130 (84-202)	0 (0-0.51)	7	0.6	4.1 (3.1-6)	87 (48-151)	0 (0-0.62)
Heart	64	3.2	1.5 (0.9-2.3)	120 (68-305)	0 (0-0.21)	41	3.3	4.0 (3.1-6.4)	69 (40-169)	0.26 (0-0.67)
Hepato-Biliary	61	3.0	0.9 (0.6-1.4)	71 (49-94)	0 (0-0.22)	39	3.1	3.6 (2.9-4.5)	53 (38-83)	0.28 (0-0.67)
Hypertension	16	0.8	1.8 (0.9-3.4)	405(97-1483)	0 (0-0.09)	17	1.4	5.3 (3.4-9.7)	118 (46-390)	0.13 (0.10-0.31)
Incontinence	28	1.4	1.4 (1.0-1.6)	67 (44-97)	0 (0-0.58)	16	1.3	3.4 (3.0-4.1)	52 (42-72)	0.16 (0-0.53)
Infectious Diseases	48	2.4	1.6 (1.0-3.3)	100 (72-339)	0.13 (0-0.50)	22	1.8	4.6 (3.5-5.3)	88 (58-145)	0.19 (0-0.61)
Inflammatory Bowel Disease and Functional Bowel Disorders	26	1.3	1.7 (1.4-2.1)	76 (59-118)	0.19 (0-0.59)	9	0.7	3.9 (3.6-4.6)	60 (54-91)	0 (0-0.15)
Injuries	30	1.5	1.3 (1.0-1.9)	78 (46-168)	0 (0-0.60)	13	1.0	3.6 (3.0-5.5)	59 (46-121)	0.31 (0.09-0.98)
Lung Cancer	6	0.3	3.5 (1.5-5.8)	233 (203-11162)	0 (0-0.08)	1	0.1	3.3 (3.3-3.3)	44	0
Menstrual Disorders and Subfertility	83	4.1	1.8 (1.2-2.3)	104 (63-144)	0 (0-0.35)	44	3.5	4.0 (3.3-4.9)	72 (46-107)	0.19 (0-0.65)
Metabolic and Endocrine Disorders	16	0.8	1.7 (0.7-2.1)	115 (67-301)	0 (0-0.16)	22	1.8	4.3 (3.0-6.5)	78 (37-175)	0.13 (0-0.49)

Table S2 (cont.) Estimated τ per Cochrane Group in the selected meta-analyses across meta-analyses

	Meta-analyses with dichotomous outcome (N=2009)					Meta-analyses with continuous outcome (N=1254)				
Cochrane Group	n	%	Study precision median (Q1-Q3)	Study N median (Q1-Q3)	Estimated τ median (Q1-Q3)	n	%	Study precision median (Q1-Q3)	Study N median (Q1-Q3)	Estimated τ median (Q1-Q3)
Movement Disorders	8	0.4	1.0 (0.7-1.3)	77 (26-150)	0 (0-0.18)	11	0.9	2.8 (2.4-5.7)	33 (22-137)	0.06 (0-0.18)
Multiple Sclerosis and Rare Diseases of the Central Nervous System	14	0.7	2.1 (1.4-2.7)	92 (71-250)	0 (0-0.31)	11	0.9	3.7 (3.5-6.2)	62 (49-157)	0.20 (0-0.3)
Musculoskeletal	39	1.9	1.4 (0.8-2.7)	73 (46-198)	0 (0-0.37)	40	3.2	3.5 (3-5.4)	57 (36-137)	0.20 (0-0.41)
Neonatal	106	5.3	1.3 (0.9-1.8)	69 (38-120)	0 (0-0.40)	66	5.3	3.8 (2.9-4.2)	57 (37-80)	0.23 (0-0.47)
Neuromuscular Disease	27	1.3	2.0 (1.3-2.5)	91 (54-220)	0 (0-0.32)	19	1.5	3.5 (2.7-6.5)	52 (35-178)	0.01 (0-0.33)
Occupational Safety and Health	5	0.2	1.5 (0.8-1.5)	92 (62-136)	0.46 (0.13-0.76)	6	0.5	4.2 (3.1-4.9)	79 (42-98)	0.15 (0-0.31)
Oral Health	27	1.3	1.2 (0.7-1.8)	51 (39-111)	0.25 (0-0.58)	20	1.6	3.7 (2.7-4.9)	54 (34-116)	0.13 (0-0.57)
Pain, Palliative and Supportive Care	71	3.5	1.9 (1.3-2.5)	106 (72-172)	0.24 (0-0.61)	29	2.3	3.6 (2.9-4.7)	54 (35-88)	0.18 (0.08-0.32)
Peripheral Vascular Diseases	37	1.8	1.5 (1-1.8)	100 (80-181)	0 (0-0.37)	21	1.7	3.8 (3.0-4.7)	60 (37-96)	0.18 (0-0.39)
Pregnancy and Childbirth	229	11.4	1.5 (1.1-2.4)	129 (77-304)	0 (0-0.44)	135	10.8	5.1 (4.0-7.9)	112 (63-260)	0.10 (0-0.37)
Prostatic Diseases and Urologic Cancers	9	0.4	1.8 (1.1-3.2)	176 (81-508)	0 (0-0.53)	6	0.5	4.3 (3.7-8.4)	83 (53-283)	0 (0-0.13)
Public Health	1	0.0	3.9 (3.9-3.9)	333	0 (0-0)	3	0.2	10.8 (5.5-11.2)	478(119-621)	0.17 (0-0.19)
Renal	43	2.1	1.1 (0.8-1.8)	72 (54-132)	0 (0-0.43)	33	2.6	3.3 (2.9-3.8)	46 (38-59)	0.26 (0-0.55)
Schizophrenia	62	3.1	1.4 (0.9-2.4)	72 (48-142)	0 (0-0.39)	35	2.8	4.8 (3.7-6.2)	104 (57-163)	0.18 (0-0.30)
Sexually Transmitted Infections	5	0.2	1.7 (1.6-2.7)	150 (68-442)	0.33 (0-0.69)	1	0.1	13.5 (13.5-13.5)	736 (736-736)	0.18 (0.18-0.18)

Table S2 (cont.) **Estimated τ per Cochrane Group in the selected meta-analyses across meta-analyses**

	Meta-analyses with dichotomous outcome (N=2009)					Meta-analyses with continuous outcome (N=1254)				
Cochrane Group	n	%	Study precision median (Q1-Q3)	Study N median (Q1-Q3)	Estimated τ median (Q1-Q3)	n	%	Study precision median (Q1-Q3)	Study N median (Q1-Q3)	Estimated τ median (Q1-Q3)
Skin	21	1.0	1.6 (1.0-1.7)	65 (39-106)	0 (0-0.15)	9	0.7	3.7 (3.3-5.2)	53 (45-113)	0.13 (0-0.38)
Stroke	62	3.1	1.6 (0.9-2.4)	95 (50-209)	0 (0-0.22)	39	3.1	3.2 (2.8-4.2)	43 (32-79)	0.17 (0-0.46)
Tobacco Addiction	39	1.9	2.4 (1.6-3.1)	263(107-524)	0.15 (0-0.33)	4	0.3	3.2(2.8-13.7)	40 (35-1177)	0 (0-0.42)
Upper Gastrointestinal and Pancreatic Diseases	34	1.7	1.2 (0.9-1.6)	88 (53-149)	0.22 (0-0.61)	15	1.2	4.4 (2.8-5.1)	77 (32-108)	0.13 (0-0.36)
Wounds	53	2.6	1.3 (1.1-1.7)	74 (50-160)	0 (0-0.42)	15	1.2	3.3 (2.3-4.6)	53 (35-84)	0.23 (0-1.01)

n: number of meta-analyses per Cochrane Group. Study precision: median and interquartile range (Q1-Q3) of the geometric means of the precision (1/SE) of the studies in a meta-analysis. Study N: median and interquartile range (Q1-Q3) of the geometric means of the numbers of subjects per study in a meta-analysis. Estimated τ: median and interquartile range (Q1-Q3).

Table S3 Estimated τ in relation to meta-analysis characteristics (n (%)) across meta-analyses

	Meta-analyses with dichotomous outcome (N=2009)							Meta-analyses with continuous outcome (N=1254)						
Characteristic	**n**	**τ=0**	**0<τ≤0.5**	**τ >0.5**	**Slope (SE)[a)]**	**p-value[a)]**	**Model R^2**	**n**	**τ=0**	**0<τ≤0.5**	**τ >0.5**	**Slope (SE)[a)]**	**p-value[a)]**	**Model R^2**
Overall	2009	1109 (55)	492 (24)	408 (20)				1254	450 (36)	513 (41)	291 (23)			
Number of studies per MA (Null model)					0.063	<0.001	0.15					0.084	<0.001	0.29
2-5	1324	790 (60)	259 (20)	275 (21)	(0.005)			927	399 (43)	336 (36)	192 (21)	(0.005)		
6-10	406	191 (47)	123 (30)	92 (23)				197	44 (22)	100 (51)	53 (27)			
>10	279	128 (46)	110 (39)	41 (15)				130	7 (5)	77 (59)	46 (35)			
Mean study size [b)]					-0.013	0.078	0.16					-0.072	<0.001	0.32
Very small	1108	731 (66)	144 (13)	233 (21)	(0.008)			282	112 (40)	68 (24)	102 (36)	(0.010)		
Small	580	255 (44)	196 (34)	129 (22)				603	209 (35)	245 (41)	149 (25)			
Medium	214	73 (34)	109 (51)	32 (15)				246	86 (35)	127 (52)	33 (13)			
Large	107	50 (47)	43 (40)	14 (13)				123	43 (35)	73 (59)	7 (6)			
Mean study N [c)]					-0.047	<0.001	0.17					-0.069	<0.001	0.32
<50	375	258 (69)	30 (8)	87 (23)	(0.007)			434	173 (40)	134 (31)	127 (29)	(0.011)		
50-99	638	365 (57)	134 (21)	139 (22)				420	145 (35)	171 (41)	104 (25)			
100-499	817	404 (49)	259 (32)	154 (19)				348	111 (32)	180 (52)	57 (16)			
≥500	179	82 (46)	69 (39)	28 (16)				52	21 (40)	28 (54)	3 (6)			

Table S3 (cont.) Estimated τ in relation to meta-analysis characteristics (n (%)) across meta-analyses

	Meta-analyses with dichotomous outcome (N=2009)							Meta-analyses with continuous outcome (N=1254)						
Characteristic	N	τ=0	0<τ≤0.5	τ >0.5	Slope (SE)[a)]	p-value[a)]	Model R^2	n	τ=0	0<τ≤0.5	τ >0.5	Slope (SE)[a)]	p-value [a)]	Model R^2
Ratio of largest/smallest study					-0.058	<0.001	0.17					-0.043	<0.001	0.30
<5	1100	659 (60)	202 (18)	239 (22)	(0.008)			826	354 (43)	292 (35)	180 (22)	(0.012)		
5-15	514	269 (52)	130 (25)	115 (22)				289	72 (25)	144 (50)	73 (25)			
>15	395	181 (46)	160 (40)	54 (14)				139	24 (17)	77 (56)	38 (27)			
MA Effect size[d)]					0.156	<0.001	0.25					0.289	<0.001	0.43
Small	481	317 (66)	128 (27)	36 (7)	(0.010)			427	212 (50)	184 (43)	31 (7)	(0.017)		
Medium	1123	602 (54)	294 (26)	227 (20)				677	213 (31)	314 (46)	150 (22)			
Large	405	190 (47)	70 (17)	145 (36)				150	25 (17)	15 (10)	110 (73)			
MA Precision [e)]					1.191	<0.001	0.54					1.816	<0.001	0.72
Very low	679	226 (33)	117 (17)	336 (49)	(0.029)			235	1 (0)	27 (11)	207 (88)	(0.042)		
Low	488	264 (54)	164 (34)	60 (12)				206	12 (6)	137 (67)	57 (28)			
Medium	326	207 (63)	109 (33)	10 (3)				241	77 (32)	144 (60)	20 (8)			
High	516	412 (80)	102 (20)	2 (0)				572	360 (63)	205 (36)	7 (1)			
Total no of subjects in MA					-0.055	<0.001	0.18					-0.065	<0.001	0.32
Small (≤500)	935	598 (64)	122 (13)	215 (23)	0.007			832	351 (42)	282 (34)	199 (24)	(0.010)		
Medium (501-1500)	539	285 (53)	145 (27)	109 (20)				260	74 (28)	131 (50)	55 (21)			
Large (>1500)	535	226 (42)	225 (42)	84 (16)				162	25 (15)	100 (62)	37 (23)			

Table S3 (cont.) **Estimated τ in relation to meta-analysis characteristics (n (%)) across meta-analyses**

	Meta-analyses with dichotomous outcome (N=2009)							Meta-analyses with continuous outcome (N=1254)						
Characteristic	N	τ=0	0<τ≤0.5	τ >0.5	Slope (SE)[a)]	p-value[a)]	Model R^2	n	τ=0	0<τ≤0.5	τ >0.5	Slope (SE)[a)]	p-value[a)]	Model R^2
Total no of events in MA					-0.027	0.002	0.16							
Small (≤150)	1129	765 (68)	152 (13)	212 (19)	(0.008)									
Medium (151-500)	522	232 (44)	160 (31)	130 (25)										
Large (>500)	358	112 (31)	180 (50)	66 (18)										
Mean event rate					0.029	<0.001	0.18							
Low (<10%)	595	437 (73)	89 (15)	69 (12)	(0.004)									
Medium (10-90%)	1384	652 (47)	399 (29)	332 (24)										
High (>90%)	30	20 (67)	4 (13)	6 (20)										

MA: meta-analysis; SE: standard error.

[a)] weighted linear regression coefficient and p-value with estimated τ (continuous) as dependent variable, versus the continuous covariates log number of studies /mean log study precision (1/SE) / $\log_{10}$ geometric mean study N/ $\log_{10}$ ratio maximum and minimum study N / absolute MA effect size (log OR for the dichotomous outcomes) / SE of effect size / $\log_{10}$ no of subjects / $\log_{10}$ no of events / log mean event rate, and adjusted for Cochrane Group and log no of studies (unless it was already in the model). [b)] Based on geometric mean study size as described in section 2.4. [c)] study N per MA. [d)] MA effect size: small effect: OR<1.2 or SMD < 0.2; medium effect: 1.2 ≤ OR ≤3 or 0.2 ≤SMD≤1; large effect: OR >3 or SMD >1; [e)] MA precision: very low: 1/SE <3; low: 3 ≤ 1/SE < 5; medium: 5 ≤ 1/SE <7.5; high: 1/SE ≥ 7.5, where SE is of the MA effect estimate.

Chapter 6

A plea for routinely presenting prediction intervals in meta-analysis

IntHout J, Ioannidis JP, Rovers MM, Goeman JJ.
Submitted

Abstract

Objectives

Evaluating the variation in the strength of the effect across studies is a key feature of meta-analyses. This variability is reflected by measures like τ^2 or I^2 [1, 2] but their clinical interpretation is not straightforward. A prediction interval is less complicated: it presents the expected range of true effects in similar studies.[3] Our objective is to recommend the prediction interval as an alternative to τ^2 or I^2.

Design

We show how the prediction interval can help understand the uncertainty about whether an intervention works or not. To evaluate the implications of the variability in the effect, we selected the first meta-analysis per intervention review of the Cochrane Database of Systematic Reviews Issues 2009-2013 with a dichotomous (n=2009) or continuous (n=1254) outcome, and compared the values in the prediction intervals with clinical thresholds.

Results

In 72.4% of 479 statistically significant (random effects $p<0.05$) meta-analyses in the Cochrane Database 2009-2013 with positive heterogeneity, the 95% prediction interval suggested that the intervention effect could be null or even be in the opposite direction. In 20.3% of those meta-analyses, the prediction interval showed that the effect could be completely opposite to the point estimate of the meta-analysis. The prediction interval can also be used to calculate the probability that a new trial will have a negative effect and to improve the calculations of the power of a new trial.

Conclusions

The prediction interval reflects the variation in treatment effects over different settings, including what effect is to be expected in future patients such as the patients that a clinician is interested to treat. Prediction intervals should be routinely reported to allow more informative inferences in meta-analyses.

Strengths and limitations of this study

- In many meta-analyses there is variation in the strength of the effect.
- The prediction interval helps in the clinical interpretation of the heterogeneity by estimating what true treatment effects can be expected in future settings.
- In case of heterogeneity, prediction intervals will show a wider range of expected treatment effects than confidence intervals, and thus may lead to different conclusions. This occurred in over 70% of statistically significant meta-analyses with positive heterogeneity of the Cochrane Database of Systematic Reviews. Completely opposite effects were not excluded in over 20% of those meta-analyses.
- Prediction intervals should be routinely reported to allow more informative inferences in meta-analyses.
- Limitations of this study are that the calculations and inferences for the prediction interval are based on the normality assumption, which may be unjustified. Further, the interval will be imprecise if the estimates of the summary effect and the τ^2 are imprecise, for example if they are based on only a few studies. Inferences based on the prediction interval are only valid for settings that are similar (exchangeable) to those on which the meta-analysis is based.

Introduction

Interventions may have heterogeneous effects across studies because of differences in study populations, interventions, follow-up length, bias, and other factors.[4] Nevertheless, the usual reporting of a meta-analysis is focused on the summary effect size combined with a confidence interval (CI) and p-value. Typically also some measure of the between-study heterogeneity is presented such as τ^2 or the inconsistency measure I^2.[1, 2] However, neither of these two metrics can readily point to the clinical implications of the observed heterogeneity. Our objective in the current article is to recommend the prediction interval as an alternative to τ^2 or I^2 because its clinical meaning is much more straightforward. The prediction interval presents the heterogeneity on the same scale as the original effect size measure, in contrast to τ^2 or I^2. Reporting a prediction interval in addition to the summary estimate and confidence interval will illustrate which range of true effects can be expected in future settings. We describe its merits and provide an example to show how it can be calculated.

Interpretation of heterogeneity

Between-study variation cannot be neglected. One of the main merits of a meta-analysis may even be that it reveals the variation of effects in different studies.[3] Therefore summarizing the findings of a meta-analysis in a single summary value sacrifices potentially informative variation.[5] However, the information that can be directly retrieved from τ^2 and I^2 with respect to the variation in the effects is limited. The clinical interpretation of I^2 is ambiguous: a high I^2 does not necessarily imply that the study effects are dispersed over a wide range[6] and a low I^2 might correspond to high dispersion[7], because I^2 depends on sample size. With very large (highly precise) studies, even tiny differences in effect size may result in a high I^2, while with small (imprecise) studies, very different treatment effects can yield an I^2 of 0. Dispersion in treatment effects is better reflected by τ because τ is the standard deviation of the between-study effects. One could for example estimate the ratio of the effect size over τ, which can convey how many times larger the treatment effect is compared to the standard deviation of the effect across studies.[8] But this may still be not very intuitive to a clinical reader. Another popular way to express variation in effect sizes is the CI, e.g. the 95% CI. The CI in a random effects model contains highly probable values for the summary treatment effect. However, it does not convey

what range of treatment effects are likely to be seen in other patients, e.g. in the next study or in the patients a clinician wants to treat in her clinic.

Prediction intervals

Not so often reported but much more insightful is the prediction interval.[9] A prediction interval always presents the heterogeneity on the same scale as the original outcomes, in contrast to τ, τ^2 or I^2. A 95% prediction interval estimates where the true effects are to be expected for 95% of similar (exchangeable) studies that might be conducted in the future.[3] Therefore it is well suited to evaluate the variability of the effect of an intervention over different settings. In the absence of between-study heterogeneity, the prediction interval coincides with the respective CI. However, in case of heterogeneity a prediction interval covers a wider range than a CI. Consequently, in case of a statistically significant effect (where all values of the 95% CI are on the same side of the null) the corresponding 95% prediction interval may indicate that values are possible on both sides of the null. This means that there will be settings where conclusions based on CIs will not hold. In the same framework, one can also calculate the probability that the true effect will be harmful (on the other side of the null) in a next study. Table 1 presents an overview of measures of between-study heterogeneity.

Table 1 Some frequently used measures for heterogeneity

Measure	Advantages	Disadvantages
τ^2		
	• τ (the square root of τ^2) is the standard deviation of the between-study variation on the scale of the original outcome • τ^2 is the direct estimate of the between-study variation and therefore useful in calculations, e.g. for the prediction interval	• A direct clinical interpretation based on τ^2 is difficult, especially when τ^2 belongs to outcomes that were analyzed on log-scale, e.g. odds ratios. • When the τ^2 estimate is based on only a few studies it will be imprecise
I^2		
	• I^2 presents the inconsistency between the study results and quantifies the proportion of observed dispersion that is real, i.e. due to between-study differences and not due to random error (Ref Higgins 2003). • I^2 reflects the extent of overlap of the confidence intervals of the study-effects. • I^2 represents the inconsistency always on a scale between 0 and 100, therefore it can be compared with suggested limits for low or high inconsistency (Cochrane handbook)	• A direct clinical interpretation of I^2 is difficult. • I^2 is also ambiguous because its size depends on sample size: o with very large studies, even tiny between-study differences in effect size may result in a high I^2 o with small (imprecise) studies, very different treatment effects can yield an I^2 of 0
Confidence interval (CI)		
	• The CI in a random effects model contains highly probable values for the summary (mean) treatment effect.	• The CI gives no information on the range of true treatment effects that are likely to be seen in other settings, e.g. in the next study or in the patients a clinician wants to treat in her clinic.
Prediction interval		
	• The prediction interval in a random effects model contains highly probable values for the true treatment effects in future settings, if those settings are similar to the settings in the meta-analysis. • The values in the interval can be compared with clinically relevant thresholds to see whether they correspond to benefit, null effects or harm. • The prediction interval can be used to estimate the probability that the treatment in a future setting will have a true positive or negative effect, and to perform better power calculations	• Conclusions drawn from the prediction interval are based on the assumption that τ^2 and the study effects are normally distributed • The estimate of the prediction interval will be imprecise if the estimates of the summary effect and the τ^2 are imprecise, for example if they are based on only a few studies and if these studies are small

Example: Topical steroids for nasal polyps

A 2012 review on the use of topical steroids for treatment of chronic rhinosinusitis with nasal polyps, based on seven randomized studies, resulted in a larger decrease in overall symptom scores in favor of steroids compared to placebo, reflected by a standardized mean difference (SMD) of -0.51, with a 95% CI from -0.96 to -0.07 (Figure 1).[10] The I^2 was 73.9% (95% CI, 44.2% to 87.8%), which can be considered substantial heterogeneity[11], and the estimated τ^2 was 0.148. Notwithstanding these numbers, it is difficult to evaluate what the clinical consequences of this heterogeneity may be for future settings.

In order to estimate the prediction interval for the SMD we need the point estimate of the SMD, its standard error (SE) and the estimated τ^2. As the SE was not reported, we derive it from the 95% CI of the SMD (formula 1 appendix), which results in an SE of 0.182. We can calculate the standard deviation of the prediction interval SD_{PI} as $\sqrt{(0.148 + 0.182^2)}$ and the lower and upper limit of the 95% prediction interval as $-0.51 \pm 2.45 \times SD_{PI}$. The value 2.45 results from the $t_{0.05/2,6}$-distribution. Prediction intervals with a different coverage could be calculated by using a different t-value, e.g. $t_{0.20/2,6}$ for an 80% prediction interval (formula 1 appendix).

The resulting prediction interval, ranging from -1.55 to 0.53, can be interpreted as the 95% range of true SMDs to be expected in similar studies. We present it in Figure 1 as a rectangle below the diamond for the 95% CI.[12] The prediction interval contains values below zero, which corresponds to a decrease in symptom scores of at best approximately 1.5 SD after steroid use compared to placebo. But it also contains values above zero which means that the steroids may exhibit no or even a harmful effect (SMD>0) in some settings, with a (95%) worst case increase in SMD of 0.53. Consequently, the effect in a new study may be even opposite to the summary point estimate of the meta-analysis, i.e. an increase of 0.51 instead of a decrease of -0.51 may occur. The estimated probability that the true effect of the steroids will be null or higher in a new study is equal to 13.6%, based on the t-distribution with 6 degrees of freedom (formula 2 appendix).

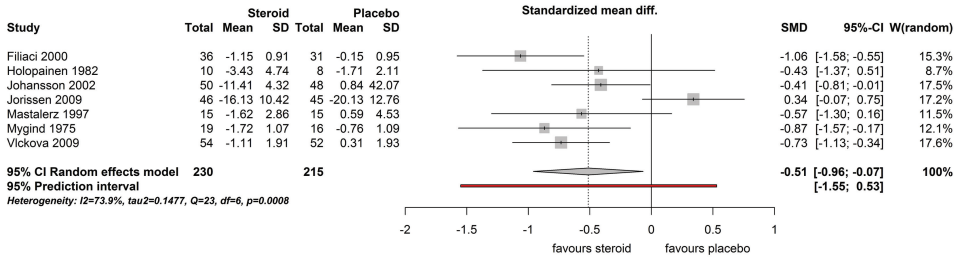

Figure 1. Forest plot of the standardized mean difference in symptom scores in nasal polyps. Steroids versus placebo, Analysis 1.1 in Cochrane Review CD006549.[10]
Note that our results differ from the original analysis, as we used a random-effects analysis with the Hartung-Knapp/Sidik-Jonkman adjustment[13] and the empirical Bayes estimator for τ^2.

Cochrane database

In order to investigate how often there is a discrepancy in conclusions based on prediction intervals and CIs we evaluated this in statistically significant meta-analyses ($p<0.05$ by random effects calculations) of the Cochrane Database of Systematic Reviews Issues 2009-2013, kindly provided by the UK Cochrane Editorial Unit. To avoid subjectivity in the selection we used the first meta-analysis with a dichotomous or continuous outcome and based on at least two studies in the Data and Analyses section. Details can be found in another paper.[14] In brief, of a total of 3263 meta-analyses, 920 were statistically significant: 479 with an estimated $I^2>0$ and 441 with an estimated $I^2=0$.

Calculations

We used the Hartung-Knapp/Sidik-Jonkman[13] random effects meta-analysis approach combined with the empirical Bayes estimator for τ^2. We estimated τ^2 for all meta-analyses, even when the authors originally performed a fixed effects analysis. Prediction intervals were calculated according to formula 1 (Appendix). We categorized the statistically significant meta-analyses with positive heterogeneity by number of studies (2-6 studies or >6) and heterogeneity ($I^2<30\%$, 30-60%, or >60%, based on the Cochrane Handbook[11] stating that an I^2 between 30% and 60% corresponds to moderate heterogeneity). For significant meta-analyses where the heterogeneity estimate was zero, we assessed the impact of possibly low but non-zero heterogeneity by assuming an I^2 of 20%, calculating prediction intervals using formula 3 (appendix). We used R software[15] version 3.1.2 and the R packages metafor[16] version 1.9-5 and meta[17] version 4.1-0.

Results

Overall, 132 (27.6%) of the 479 statistically significant meta-analyses with an $I^2 > 0$ had both the 95% CI and the 95% prediction interval excluding the null effect (Table 2). Consequently, almost three-quarter (72.4%) had a prediction interval that contained the null effect. This means that it is likely that for these comparisons some patient populations might experience null effects or effects in the opposite direction, i.e. a treatment might be more harmful than the comparator even though the point estimate suggests benefit (or vice versa). Not surprisingly, significant meta-analyses with low heterogeneity more often had prediction intervals that excluded the null than meta-analyses with high

heterogeneity. The percentage of prediction intervals containing the null effect was slightly higher for meta-analyses with a continuous outcome (80.4%) than for those with a dichotomous outcome (65.8%), and for meta-analyses based on more than six studies (74.1%) than for those with at most six studies (69.1%). (Tables S1.A and S1.B in the Appendix).

Prediction intervals containing the opposite effect

If the prediction interval just includes the null effect, this may be less worrying than when it contains the opposite effect of the pooled summary effect, e.g. if it contains an OR of 0.5 when the meta-analysis summary estimate is an OR of 2, or if it contains an SMD of -0.7 when the summary estimate was 0.7. Of the 479 significant meta-analyses with an $I^2 > 0$, 97 (20.3%) had a prediction interval that contained the opposite effect. This percentage was higher for the meta-analyses with a continuous outcome (65/219, 29.7%) than for those with a dichotomous outcome (32/260, 12.3%). It occurred also more frequently in meta-analyses with more than 6 primary studies (57/139 and 30/178 for meta-analyses with a continuous or dichotomous outcome, respectively) than for those based on at most 6 studies (8/80 and 2/82).

Meta-analyses with estimated $I^2 = 0$

A substantial part of meta-analyses have an estimated I^2 of 0. However, there is typically very large uncertainty about the exact amount of heterogeneity and this is demonstrated by very large 95% CIs for the values of I^2.[18] The same applies to τ: an estimate of 0 is often accompanied by large uncertainty. The true I^2 and τ are unlikely to ever be exactly 0, although low values are possible. To assess the impact of possibly low but non-zero heterogeneity among the 441 Cochrane meta-analyses with estimated $I^2=0$ and statistically significant results, we imputed an $I^2=20\%$ (suggestive of low between-study heterogeneity). Under this assumption, in 329 (74.6%) of these 441 meta-analyses the 95% prediction interval would span both sides of the null (Table 2), similar for meta-analyses with a dichotomous (74.7%) or continuous (74.4%) outcome (Tables S1.A in the Appendix). This is a sensitivity analysis that is useful to perform to see whether the inferences of a meta-analysis that seemingly does not have detectable heterogeneity may be influenced by even a small amount of heterogeneity.

Table 2 Proportion of statistically significant meta-analyses where both the 95% confidence and prediction intervals excluded the null

	Estimated heterogeneity I^2				
Statistically significant meta-analyses	**I^2=0[a)]**	**I^2>0**	**>0–30%**	**30-60%**	**>60%**
N	441	479	123	150	206
Both 95% CI and 95% PI excluded null (n (%))	112 (25.4)	132 (27.6)	88 (71.5)	39 (26.0)	5 (2.4)

CI: confidence interval; PI: prediction interval.

[a)] When the estimated heterogeneity I^2 was equal to 0, I^2=20% was imputed for the calculation of the prediction interval.

Power calculations for a future study

Meta-analysis results can also be used for power calculations for a new study. However, the expected true effect in a new study is not necessarily equal to the point estimate of the meta-analysis: it can be any of the values in the prediction interval. In case of heterogeneity an apparent power of 80% based on the point estimate will be overly optimistic because the power function is asymmetric. If the true study effect is larger than the point estimate the real power of the study will be higher, up to a maximum of 100%, but if the effect is smaller the power may decrease substantially, even to 5% or less in case of a null effect. Consequently the expected power of a new study in case of heterogeneity will be lower than 80% (formula 4 appendix). For example, if the prediction interval shows that 30% of future studies may have a true null or negative effect, the power can never be much larger than 70%. The sample size should be increased to compensate for this loss in power, see also Roloff et al.[19]

Conclusions and outlook

In meta-analyses a CI is inadequate for clinical decision making because it only summarizes the average effect for the average study. The prediction interval is more informative as it shows the range of possible effects in relation to harm and clinical benefit thresholds. While we have focused on the situation where the separating threshold is the null, a different threshold may be considered. For example, in the prediction interval framework one can calculate the probability that an effect is larger than B, where B may be a clinically meaningful effect (if the treatment benefit is less than B, then it is felt not to be worth it). A narrow prediction interval that lies completely on the beneficial side of a clinically relevant threshold increases confidence in an intervention. A broad prediction interval may indicate the existence of settings where the treatment has a suboptimal and possibly even harmful effect. In more than 70% of statistically significant meta-analyses of the Cochrane Database with some estimated or assumed between-study heterogeneity the prediction intervals crossed the no-effect threshold, indicating that there are settings where those treatments will have no effect or even an effect in the opposite direction. In 20.3% of those meta-analyses the prediction interval even contained the opposite effect of the summary estimate, for example an OR of 0.5 when the summary point estimate was an OR of 2. This occurred most frequently for meta-analyses with a continuous outcome and for meta-analyses based on more than six studies, probably because such meta-analyses have more power to detect smaller effects, which means that also the opposite effects will be smaller.

Graham en Moran[20] evaluated prediction intervals in 72 meta-analyses with a dichotomous outcome in critical care published between 2002 and 2010. They found a higher percentage of significant meta-analyses (50/72, 69.4%), compared to 28.5% (572/2009) in our set of meta-analyses with an odds ratio outcome. The difference may be caused by publication bias, the higher number of primary studies in their sample (medium 9 versus 4 in our set[14]), and by their use of the DerSimonian-Laird approach which can result in too many significant findings, whereas we used the HKSJ approach.[13] However, results with respect to the prediction interval were remarkably similar. In 32 (64.0%) of their 50 significant meta-analyses the 95% prediction interval included the null, similar to 65.8% in our dataset. Seven (14.0%) of their 50 meta-analyses suggested a high probability

of efficacy or harm, similar to 12.3% of our meta-analyses where the prediction interval contained the opposite effect, despite the fact that they used a different definitions for possible “harm” and that they do not mention whether there was positive between-study heterogeneity in their significant meta-analyses.

It is straightforward to calculate a prediction interval if we can assume that the effects are normally distributed and that τ^2 is known and stable across studies. However, one should realize that the prediction interval is dependent on this assumption and on the precisions of the estimated τ^2 and study effect, and will be imprecise if the number of studies in the meta-analysis is small. If the number of studies is large, estimates will be more precise and the normality of the distribution of τ^2 can be evaluated. A final caveat is that the uncertainty conveyed by the prediction interval pertains to the uncertainty about the extent to which future studies are similar (exchangeable) to those that have already been done, but this applies to all inferences from a meta-analysis. If the future studies evaluate patients and settings that are entirely different from what was evaluated in past studies, this exchangeability is questionable and uncertainty may be even more prominent than what the prediction interval conveys. In practical terms, if the patients treated by a physician are considered to be very different from the patients seen in all studies that have been done in the past, even the prediction interval cannot tell us what we might expect for these patients.

Summarizing, the prediction interval reflects the variation in true treatment effects over different settings, including what effect is to be expected in future patients such as the patients that a clinician is interested to treat. Therefore it should be routinely reported in addition to the summary effect and its confidence interval, and used as a main tool for interpreting evidence, to enable more informed clinical decision making.

Appendix

Formula 1 Prediction interval

In order to calculate the 95% prediction interval, the summary meta-analysis estimate M, the two sided critical t-value $t_{0.05/2,\ k-1}$ and the standard deviation for the prediction interval SD_{PI} are needed. Here, *t* is the two-sided critical t-value that can be calculated via http://www.danielsoper.com/statcalc3/calc.aspx?id=10. Fill in DF=k-1 and probability level 0.025, with *k* the number of studies in the meta-analysis. SD_{PI} is the standard deviation of the prediction interval: $SD_{PI} = \sqrt{(\tau^2+SE^2)}$, where τ^2 is the estimated heterogeneity and SE is the standard error of M[4, 16]. If the SE was not reported, it can be approximated by dividing the distance between the limits of the 95% CI of the SMD by 3.92. The lower and upper limits of the 95% prediction interval are equal to $M \pm t_{0.05/2,\ k-1} \times SD_{PI}$. Of course it is possible to estimate prediction intervals with a different coverage, e.g. an 80% prediction interval would be based on $t_{0.20/2,6}$.

Estimations for ORs, risk ratios and hazard ratios are generally performed on the natural logarithm scale. As an example we take the calculation of a 95% prediction interval for an OR of 2.28 with a 95% CI from 1.05 to 4.96, $\tau^2 = 0.353$ and *k*=7. The prediction interval will first be estimated on log scale. Note that the reported τ^2 is in general already the heterogeneity for log OR, not for OR, and can thus be used directly in the calculations. The SE of the log OR is calculated by dividing the distance between the log of the limits of the 95% CI of the OR by 3.92. This results in SE=0.318.

The lower and upper limits of the 95% prediction interval for the log OR are $\log(2.28) \pm 2.45\sqrt{(0.353 + 0.318^2)}$. The value 2.45 results from the $t_{0.05/2}$-distribution with 6 DF. Finally, we exponentiate the limits to return to the OR scale. The resulting prediction interval ranges from 0.44 to 11.86, and can be interpreted as the 95% range of true ORs to be expected in similar studies.

Formula 2 Probability that effect is larger than threshold D

The probability P that the true effect in a new study will be below a threshold D (e.g. the null effect) can be calculated with the left-tail cumulative t-distribution with k-1 degrees of freedom. The probability that the effect is above D equals 1 – P.

In our example on nasal polyps the probability that the SMD ≥ 0 can be estimated as follows:

1. Start to calculate the probability P that a true SMD ≤ 0. This is equivalent to the probability that a t-value ≤ T, where T is equal to $(D - M)/SD_{PI}$, with summary treatment effect M= -0.51, SD_{PI} = 0.425 and D=0. This results in T = 1.207, with 6 degrees of freedom (DF).
2. The probability P can be calculated online at http://www.danielsoper.com/statcalc3/calc.aspx?id=41. Fill in t value = 1.207 and DF = 6. The one-tailed probability P(t ≤1.207)= 0.864.
3. We want the probability that the SMD ≥0, this is 1 – P= 0.136.

In the example on the OR (see formula 1), if we are interested in the probability of a null or negative effect, we are interested in the probability that a true OR ≤ 1. For ORs, calculations must be based on the ln OR, with M= ln(2.28)=0.824, SD_{PI} = 0.674, and DF=6. A true OR ≤ 1 corresponds to a true ln OR ≤ 0.
Fill in T = (0-0.824)/0.674 = -1.223 and DF=6. The probability that a true OR ≤ 1 is equal to 0.134.

Formula 3 Prediction interval starting with I^2

In order to calculate prediction intervals starting with an assumed I^2 value (as percentage), we first calculated the corresponding τ^2value:

$$\tau^2 = s^2 \frac{I^2}{100 - I^2}$$

with s^2 the typical study variance, equal to $\frac{\sum w_i (k-1)}{(\sum W_i)^2 - \sum w_i^2}$,
and w_i equal to the inverse of the study variance of study i (i=1..k) and k the number of studies.[21] Subsequently formula 1 can be applied.

Formula 4 Power of a future study

Usually sample size calculations are performed without consideration of the heterogeneity. If we do take into account the heterogeneity, the expected power, i.e. the probability that a new study with N patients will have a positive result at significance level α, given values for the standard error *s* of the new study and μ and τ^2 as above, can be approximated with the delta method if τ^2 is not too large:

$$E(power) = g(\mu) + 0.5\,\tau^2\,g''(\mu)$$

where g is the power at the meta-analysis summary estimate μ, and $g''(\mu)$ is the second derivative of g at μ. For $g''(\mu)$ we can take the second derivative of the normal cumulative distribution function if N is sufficiently large.

This results in $g''(\mu) = \frac{z_\mu e^{-0.5z_\mu^2}}{s^2\sqrt{2\pi}}$, with $z_\mu = \frac{1.96s - \mu}{s}$.

If the sample size N of the new study is such that the power for an effect of size μ is 80%, the expected power of the study will be smaller than 80% if τ^2 is positive, because the corresponding value of z_μ is negative.

Table S1.A **Proportion of statistically significant meta-analyses where both the 95% confidence and 95% prediction intervals excluded the null Separately for dichotomous and continuous outcomes**

	I^2=0	I^2>0	<30	30-60	>60
All meta-analyses (N=3263)					
MA stat. significant (N)	441	479	123	150	206
Both 95% CI and 95% PI excluded the null [a)]	112	132	88	39	5
(N (%))	(25.4)	(27.6)	(71.5)	(26.0)	(2.4)
MAs with dichotomous outcome (N=2009)					
MA stat. significant (N)	312	260	88	96	76
Both 95% CI and 95% PI excluded the null [a)]	79	89	61	23	5
(N (%))	(25.3)	(34.2)	(69.3)	(24.0)	(6.6)
MAs with continuous outcome (N=1254)					
MA stat. significant (N)	129	219	35	54	130
Both 95% CI and 95% PI excluded the null [a)]	33	43	27	16	0
(N (%))	(25.6)	(19.6)	(77.1)	(29.6)	(0.0)

MA: meta-analysis; CI= 95% confidence interval; PI= 95% prediction interval;
[a)] When the estimated heterogeneity I^2 was equal to 0, I^2=20% was imputed for the calculation of the prediction interval

Table S1.B **Proportion of statistically significant meta-analyses where both the 95% confidence and 95% prediction intervals excluded the null Separately for dichotomous and continuous outcomes and 2-6 vs. >6 studies**

2-6 studies	**I^2=0**	**<30**	**30-60**	**>60**
All meta-analyses				
MA stat. significant (N)	322	44	59	59
Both 95% CI and 95% PI excluded the null [a)] (N (%))	74 (23.0)	32 (77.7)	17 (18.8)	1 (1.7)
MAs with dichotomous outcome				
MA stat. significant (N)	210	32	30	20
Both 95% CI and 95% PI excluded the null [a)] (N (%))	50 (23.8)	24 (75.0)	7 (23.3)	1 (5.0)
MAs with continuous outcome				
MA stat. significant (N)	112	12	29	39
Both 95% CI and 95% PI excluded the null [a)] (N (%))	24 (21.4)	8 (66.7)	10 (34.5)	0 (0.0)
>6 studies	**I^2=0**	**<30**	**30-60**	**>60**
All meta-analyses				
MA stat. significant (N)	119	79	91	147
Both 95% CI and 95% PI excluded the null [a)] (N (%))	38 (31.9)	56 (70.9)	22 (24.2)	4 2.7)
MAs with dichotomous outcome				
MA stat. significant (N)	102	56	66	56
Both 95% CI and 95% PI excluded the null [a)] (N (%))	29 (28.4)	37 (66.1)	16 (24.2)	4 (7.1)
MAs with continuous outcome				
MA stat. significant (N)	17	23	25	91
Both 95% CI and 95% PI excluded the null [a)] (N (%))	9 (52.9)	19 (82.6)	6 (24.0)	0 (0.0)

MA: meta-analysis; CI= 95% confidence interval; PI= 95% prediction interval;
[a)] When the estimated heterogeneity I^2 was equal to 0, I^2=20% was imputed for the calculation of the prediction interval

References

1. Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? Statistics in Medicine. 2002;21(11):1559-73.
2. Higgins J, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003;327(7414):557.
3. Higgins J, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2009;172(1):137-59.
4. Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. BMJ. 2011;342.
5. Saha S, Chant D, McGrath J. Meta-analyses of the incidence and prevalence of schizophrenia: conceptual and methodological issues. International Journal of Methods in Psychiatric Research. 2008;17(1):55-61.
6. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to Meta-Analysis. Chichester, UK: Wiley; 2009.
7. Melsen WG, Bootsma MCJ, Rovers MM, Bonten MJM. The effects of clinical and statistical heterogeneity on the predictive values of results from meta-analyses. Clinical Microbiology and Infection. 2014;20(2):123-9.
8. Moonesinghe R, Khoury MJ, Liu T, Ioannidis JPA. Required sample size and nonreplicability thresholds for heterogeneous genetic associations. Proceedings of the National Academy of Sciences. 2008;105(2):617-22.
9. Chiolero A, Santschi V, Burnand B, Platt R, Paradis G. Meta-analyses: with confidence or prediction intervals? European Journal of Epidemiology. 2012;27(10):823-5.
10. Kalish L, Snidvongs K, Sivasubramaniam R, Cope D, Harvey RJ. Topical steroids for nasal polyps. Cochrane Database Syst Rev. 2012;12:CD006549.
11. Higgins JPT, Green S, Collaboration C. Cochrane handbook for systematic reviews of interventions: Wiley Online Library; 2008.
12. Guddat C, Grouven U, Bender R, Skipka G. A note on the graphical presentation of prediction intervals in random-effects meta-analyses. Systematic Reviews. 2012;1(1):34.
13. IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC medical research methodology. 2014;14(1):25.

14. IntHout J, Ioannidis JPA, Borm GF, Goeman JJ. Small studies are more heterogeneous than large ones: a meta-meta-analysis. Journal of Clinical Epidemiology. 2015;68(8):860-9.
15. R Core Team. R: A language and environment for statistical computing. Retrieved from http://www.R-project.org/. Vienna, Austria: R Foundation for Statistical Computing; 2014.
16. Viechtbauer W. Conducting meta-analyses in R with the metafor package. Journal of Statistical Software. 2010;36(3):1-48.
17. Schwarzer G. meta: General Package for Meta-Analysis. R package version 4.1-0. http://CRAN.R-project.org/package=meta. 2015.
18. Ioannidis JPA, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analyses. BMJ. 2007;335(7626):914-6.
19. Roloff V, Higgins J, Sutton AJ. Planning future studies based on the conditional power of a meta-analysis. Statistics in Medicine. 2013;32(1):11-24.
20. Graham PL, Moran JL. Robust meta-analytic conclusions mandate the provision of prediction intervals in meta-analysis summaries. Journal of Clinical Epidemiology. 2012;65(5):503-10.
21. Bowden J, Tierney JF, Copas AJ, Burdett S. Quantifying, displaying and accounting for heterogeneity in the meta-analysis of RCTs using standard and generalised Q statistics. BMC Medical Research Methodology. 2011;11(1):41.

Chapter 7

Discussion

The aim of this thesis was to give more insight into the application of meta-analysis methodology and to reflect on the role of between-study heterogeneity in the realistic setting where most meta-analyses are based on just a few studies and where some of these studies are small or very small.

There is much interest in the application of meta-analysis, as it provides a suitable quantitative method to summarize the continuously increasing amount of information resulting from more than 75 publications on randomized controlled trials per day.[1] Moreover, a meta-analysis can result in a better estimate of an effect size than a new large study, particularly if there is between-study variation in the results, which is the case in approximately 50% of the meta-analyses. Due to this heterogeneity the results of a single trial may be overly positive or negative, and one cannot know how far they deviate from the average effect without knowing the results of other studies. The simulations presented in **Chapter 3** show that even if there are only two or three primary studies - possibly subject to publication bias - the estimate of the effect size provided by the meta-analysis is preferable to the results of a single new trial, because the meta-analysis results in lower error rates for false positive findings.

Estimates of the between-study heterogeneity are imprecise if they are based on small numbers of studies. Data retrieved from the Cochrane Database of Systematic Reviews Issues 2009-2013 clearly show that many meta-analyses are based on only a few (medium 4, interquartile range 2-6) primary studies. The imprecision of the estimated heterogeneity is neglected in the commonly used DerSimonian and Laird (DL) method for random effects meta-analysis. For meta-analyses with realistic numbers of primary studies (2-20 studies) we showed that a simple variation on the DL method, developed by Hartung and Knapp and also by Sidik & Jonkman (HKSJ) results in more adequate false positive error rates than the DL method (**Chapter 4**). Of course there exist other, possibly more advanced methods that result in an improvement of error rates compared to DL, for example the restricted maximum likelihood method. However, the HKSJ method has been shown to perform very well in simulations[2] and is easy to apply.

In **Chapter 5** we showed that the estimated heterogeneity between small studies was larger than between large studies and its imprecision was also larger. Apparently results of small studies are more different from each other than those of large studies, even after adjustment for the additional random variation due to the smaller sample sizes. The common weighting of studies in the random effects approach is based on only one estimate for the heterogeneity, equal for small and large studies. Consequently, results of small studies will receive too much emphasis in the summary estimate. However, the low number of primary studies makes it difficult to provide a single reliable estimate of the between study heterogeneity in most meta-analyses; it will be even harder to provide study size dependent estimates. Another complicating factor is that studies that are selected for a meta-analysis are not a random sample of all possible studies, but are conducted in a setting where one study may be a reaction on another study. For example, after a small study with promising results a number of larger studies may be started. However, if the outcome of a first study was disappointing or counterintuitive, only a small trial may be started or the research might be discontinued. A basic assumption of the random effects approach is that the primary studies are a random sample of all possible studies, and that results may be generalized to similar settings or studies. Clearly further research in this area is needed. Henmi et al.[3] advocated the combination of a fixed effect estimate with a confidence interval based on a random effects model. Although their objective was to make results of meta-analyses less sensitive to the effects of publication bias, approaches like this might be a way to proceed if small studies seem too influential

If there is heterogeneity, i.e. when effect sizes differ widely over the studies, this variation is clinically relevant and should be clearly visible in the results. For example if some results strongly favour the experimental treatment whereas others strongly favour the control, this is relevant information for the interpretation of the results of the meta-analysis. Usually the extent of heterogeneity is reported with an estimate of τ^2 or I^2, in addition to the summary estimate of the meta-analysis and its confidence interval. However, we argue that this approach is too limited. A prediction interval shows the expected range of true effect sizes in future settings, whereas a confidence interval only represents the summary effect size and its imprecision. In **Chapter 6** we wrote a plea for routinely presenting the prediction interval as a result of meta-analyses, in addition to the confidence interval and the summary estimate. More than τ^2 or

I^2, the prediction interval can help the reader to understand the uncertainty about whether a treatment works or not. In roughly 70% of statistically significant (random effects $p<0.05$) meta-analyses in the CDSR Issues 2009-2013, the 95% prediction interval suggested that the intervention effect could even be in the opposite direction, and in approximately 20% of these meta-analyses the opposite effect was as least as large as the summary estimate of the meta-analysis (e.g. the prediction interval contained an OR of 0.5 when the point estimate was an OR of 2. Although we were not the first to describe the advantages of presenting prediction intervals, it is still far from common practice in meta-analyses of interventions, in contrast to prediction regions in meta-analyses on diagnostic test accuracy. Perhaps this is related to the fact that the methodology in diagnostic test accuracy meta-analyses in general is more complex and further developed than in the traditional pairwise meta-analyses, as the correlation between sensitivity, specificity and prevalence has to be incorporated in the statistical models.

Although meta-analyses on animal studies are far less often performed than meta-analyses on clinical studies, the advantages of a meta-analysis on animal studies may be even larger. Results of animal studies are more heterogeneous than those of human studies, therefore the results of a meta-analysis can be more reliable than those of a new trial. Besides less animals need to be sacrificed. Further, we hope that as a consequence of the increased emphasis on "old" study results, quality of reporting will improve. Therefore we wrote an introductory tutorial on meta-analysis methods for animal studies (**Chapter 2**). However, animal studies have their complexities, even more than human studies. First, they are extremely small compared to most human studies. Their results show high heterogeneity: responses of 0% and 100% in the same meta-analysis are far from uncommon. Designs may include measurements over time, resulting from a few animals that are sacrificed each time period. The same control group may be used repeatedly. Moreover there is still ample room for improvement in the quality of reporting of the studies. The nature of the tutorial was explicitly introductory, aiming at animal scientists who want to learn more about meta-analyses. There is ample room for a more advanced tutorial focussing on more sophisticated statistical analysis methods, as shown by the list of complexities specific to animal studies. However, this requires additional methodological research.

Outlook

In 1959 it was already stated that approaches to meta-analysis were too simple and that more advanced methodology should be applied.[4] In 2015 Hoaglin repeated this complaint and stated that, despite the extensive literature, most of the methodology development had limited relevance to actual practice.[5] As reasons he mentioned the use of incorrect (but convenient) assumptions, inaccurate approximations, preference for simplicity, unrealistic simulations and inertia. However, recently there have been quite a few developments. Up to a few years ago meta-analyses were usually paired comparisons leading to summary estimates based on inverse variance approach, with as major discussion point whether a random or a fixed effects model was to be used. Nowadays the share of network meta-analyses, individual participant meta-analyses and diagnostic test accuracy meta-analyses is increasing. Also a more advanced method like the multivariate approach is gaining popularity, since Van Houwelingen et al. wrote a tutorial in 2002.[6] Multivariate techniques are the standard approach already for network and diagnostic meta-analyses. However, they can be rather complicated and it will be difficult to turn them into automated procedures like Review Manager (RevMan).[7] Another advanced method is the Bayesian approach, which explicitly models the imprecision. The Bayesian approach also offers opportunities, like a ranking of interventions in a network meta-analysis based on the probability that an intervention is the most efficacious one, that are difficult to perform with a frequentist analysis. It was even suggested to analyse single studies from a Bayesian meta-analysis perspective, in order to shrink unrealistically extreme study effect estimates towards the prior mean.[8] However, the Bayesian approach is still far from mainstream, even though statistical software to implement it is more readily available since it has been incorporated in some SAS procedures and R packages.

Summarizing, the landscape of evidence based medicine is changing, and the methodology of meta-analysis is maturing. The evidence base is changing from the single study results via paired treatment comparisons based on a group of studies, to a set of studies evaluating similar interventions in one large (network) meta-analysis. There are many initiatives stimulating the data sharing of individual participant data. The availability of individual data will allow a consistent approach to the decisions that must be taken during the preparations for the analysis and the statistical models. This will enable a more consistent

approach to the meta-analysis as well as more thorough insight in the reasons for the heterogeneity. Currently, most source data consist of aggregated study data, which makes it difficult to deal with heterogeneity. If evident and measurable clinical reasons for heterogeneity exist, they should be taken into consideration and investigated. It is possible to use meta-regression to adjust for differences between studies and to possibly decrease the extent of residual between-study heterogeneity. However, especially for variables that summarize individual patient characteristics on study level, like mean age, results of a meta-regression may be unreliable due to the ecological fallacy[9], that may result in an apparent relationship between the moderator variable and the treatment effect that may even be opposite to the one that would be found if individual patient data would have been used. Another complication is that there will be studies that did not report aggregated data for the moderator variable. Up till now, not much research has been published on missing data imputation for meta-regression. However, at least part of these problems will be solved once individual participant data will be available.

Multivariate methods are available to answer advanced research questions in network meta-analysis or individual participant meta-analyses. These methods enable also the incorporation of more complex designs of the primary studies, e.g. stepped wedge or repeated measures designs. In addition to outcomes based on approximations of the normal distribution, also outcomes that have a binomial or other distribution can be modelled. Missing data imputation is another opportunity for methodological development.

Recommendations

Heterogeneity, either clinical or statistical, is of direct importance for the size and the interpretation of the results of the meta-analysis. First of all, the meta-analyst should take care that it is not caused by careless study selection. Unsuitable studies (for example with a too different target population) should not be included in meta-analysis.

Due to the limited number of primary studies in most meta-analyses, the estimate of the heterogeneity may be imprecise. We recommend that the meta-analysis method should recognize this fact, and not be too reliant on large sample approximations where possible. Ideally, the method should be robust, and

insensitive to the exact size of the estimated heterogeneity. Sensitivity analyses may be performed by using some values from the confidence interval of the τ^2 estimate.

Further we recommend that the heterogeneity estimate should be presented with prediction intervals on the scale of the clinical outcome, so that results of a meta-analysis illustrate which range of true effects can be expected in future settings, for example in a hospital. Meta-analysis based on individual participant data will greatly increase the possibility to get insight in the reasons for the heterogeneity, which is very important for the interpretation of the variation in the study results. Whenever possible, the reasons and the remaining unexplained heterogeneity should be clearly reported, so that individual doctors may assess how large the effect sizes are that they may expect if they apply the intervention in their hospital.

Conclusion

The underlying rationale for the use of a random-effects approach to meta-analysis is the acceptance of heterogeneity in effect sizes. Heterogeneity plays an important role both in the statistical estimation of the summary effect and the clinical interpretation of the results of a meta-analysis. Evaluating the variation in the strength of the effect across studies is a key feature of meta-analyses.

References

1. Bastian H, Glasziou P, Chalmers I. Seventy-Five Trials and Eleven Systematic Reviews a Day: How Will We Ever Keep Up? PLoS Med. 2010;7(9):e1000326.
2. Sánchez-Meca J, Marín-Martínez F. Confidence intervals for the overall effect size in random-effects meta-analysis. Psychological methods. 2008;13(1):31.
3. Henmi M, Copas JB. Confidence intervals for random effects meta-analysis and robustness to publication bias. Statistics in Medicine. 2010;29(29):2969-83.
4. Dickersin K. Innovation and cross-fertilization in systematic reviews and meta-analysis: The influence of women investigators. Research Synthesis Methods. 2015.
5. Hoaglin DC. We know less than we should about methods of meta-analysis. Research Synthesis Methods. 2015.
6. Van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Statistics in Medicine. 2002;21(4):589-624.
7. The Cochrane Collaboration. Review Manager (RevMan) 5.1.4. Copenhagen: The Nordic Cochrane Centre; 2011.
8. Gelman A. Working through some issues. Significance. 2015;12(3):33-5.
9. Da Costa BR, Jüni P. Systematic reviews and meta-analyses of randomized trials: principles and pitfalls. European Heart Journal. 2014;35:3336–45.

Chapter 8

Summary
Nederlandse samenvatting

Summary

The theme of the Cochrane Colloquium 2015 is “Filtering the information overload for better health decisions”. Per day more than 75 publications on randomized controlled trials are published; a huge amount of information, which is even continuously increasing. One of the most important quantitative tools which can be used to extract valuable information from this data overload is meta-analysis.

Study reproducibility is another issue which is high on the research agenda. Study results are difficult to reproduce, even of studies that are published in high-impact journals. Publication bias and variation due to sample size limitations were mentioned as possible reasons. However, another reason may be the between-study variation in the results caused by clinical or statistical differences between studies.

The main objective of this thesis is to give insight into the application of meta-analysis methodology and to reflect on the role of between-study heterogeneity in the realistic setting where most meta-analyses are based on just a few studies and where some of these studies are small or very small. Default meta-analysis methodology is based on large sample approximations, for which a sufficient number of studies must be available.

The UK Cochrane Editorial Unit kindly provided us with the statistical data of systematic reviews of interventions in clinical studies included in the Cochrane Database of Systematic Reviews (CDSR) Issues of 2009-2013. We used these data to empirically evaluate various aspects of meta-analysis methodology and to design realistic simulation studies for more extensive evaluations.

Meta-analyses of animal studies – tutorial

Most medical meta-analyses are dedicated to clinical (human) research. However, some questions can only be answered by means of experimental animal studies. The number of meta-analyses of animal studies is increasing, but it is still very much lower than the number of meta-analyses of clinical studies, even though the statistical methodology is rather similar. The advantages of a meta-analysis on animal studies may be even larger than for clinical studies. Results of animal studies are more heterogeneous than those of human studies, as animal studies are often more diverse in their populations (e.g., species), design, and other study characteristics. Therefore the pooled results of a meta-analysis can be more reliable than those of a new trial. If the meta-analysis provides a convincing result, fewer animals need to be sacrificed. Further, the increased emphasis on "old" study results will reveal gaps in the reporting of those results. We hope that as a consequence reporting quality will improve. To explain the methodology and to stimulate animal researchers to perform more meta-analyses, we have written an introductory tutorial, which is presented in **Chapter 2**. However, animal studies have their complexities, even more than human studies: they can be very small, show high between-study heterogeneity in effect sizes, use a variety of designs, and the quality of reporting needs improvement. There is ample room for a more advanced tutorial focussing on more sophisticated statistical analysis methods. However, this requires additional methodological research.

Meta-analysis of several modestly powered trials outperformed the evidence generated by a single well-powered trial

In **Chapter 3** we investigate a fundamental question with regard to two possible approaches to find the best evidence on the effect of an intervention: is it preferable to conduct one large new trial or it is better to summarize existing trial results by means of a meta-analysis. We also evaluate the effect of three complicating factors. First, the number and size of the trials: if a meta-analysis is preferred to a large trial, is this then also the case if only a few small trials are available for the meta-analysis? Secondly, we investigate the influence of reporting bias. Studies with positive findings tend to get published more often than those with negative findings, which results in an overrepresentation of papers with positive results. The third issue is the between study variation, or heterogeneity.

Extensive simulations show that a meta-analysis tended to result in a better estimate of an effect size than a new large study, particularly if there was between-study variation in the treatment effect, which is shown to be the case in approximately 50% of the meta-analyses. Due to this heterogeneity the results of a single trial may be overly positive or negative, and one cannot know how far they deviate from the average effect without knowing the results of other studies. When a treatment was assumed to have no effect but heterogeneity was present, the error rates for a single trial were increased more than tenfold above the nominal rate, even for low heterogeneity. In contrast, for meta-analyses on series of trials the error rates were correct. When selective publication was present, the error rates were always increased, but they still tended to be lower for series of trials than for single trials. Even if there were only two or three primary studies, subject to publication bias and heterogeneity, the meta-analysis resulted in lower rates of false positive findings than a single large new trial.

The Hartung-Knapp-Sidik-Jonkman method outperformed the standard DerSimonian-Laird method

In a random effects meta-analysis the size of the estimated heterogeneity directly affects the meta-analysis result through the weighting of the primary studies. If a meta-analysis is based on only a few studies, the estimate of the heterogeneity will be imprecise. Data retrieved from the CDSR Issues 2009-2013 show that many meta-analyses were based on only a few (medium 4, interquartile range 2-6) primary studies. However, the default method for random effects meta-analysis is (still) the DerSimonian and Laird (DL) approach, even though it is based on large sample approximations and the imprecision of the estimated heterogeneity is neglected. The method described by Hartung, Knapp(11-13) and by Sidik and Jonkman(14, 15) (HKSJ) is known to perform better, but evidence in realistic situations, where one trial might be much larger than the other trials, was lacking. In **Chapter 4** we assessed the relative performance of those methods when studies of different sizes are combined. We simulated meta-analyses of 2–20 trials with varying sample sizes and between-study heterogeneity. Results show that the HKSJ method resulted in lower, more adequate false positive error rates than the DL method. We illustrated this empirically by demonstrating that the number of “positive” (statistically significant at $p < 0.05$) findings using the DL approach was much higher than by the HKSJ approach, using 689 meta-analyses from the CDSR Issues 2012. Further

we showed how DL results can be converted to HKSJ results by means of a few steps and provided an Excel sheet for this conversion.

Small studies were more heterogeneous than large ones

Results of small studies are often associated with "small study effects": the phenomenon that results of small studies tend to be more positive than those of larger studies. This can be due to reporting bias but also to quality issues, both of which may occur more often in small trials. Whether also the heterogeneity between small studies is different from the heterogeneity between large studies, was investigated in **Chapter 5.** By exploring 3263 meta-analyses from the CDSR Issues 2009-2013 we showed that the estimated heterogeneity between small studies was larger than between large studies and its imprecision was also larger. Results of small studies are more different from each other than those of large studies, even after adjustment for the additional random variation due to the smaller sample sizes.

A plea for routinely presenting prediction intervals in meta-analysis

Heterogeneity is not only associated to the weights of the studies in the meta-analysis: it also contains clinically relevant information. The existence of a positive heterogeneity estimate implies that there may be differences in the intervention effect between the trials. This occurs often: in approximately half of the meta-analyses the estimate for the between-study variation is positive. This means that the treatment will appear more effective in some settings than in others, which clearly is clinically relevant. However, most reviewers and readers are uncertain with respect to the clinical interpretation of the heterogeneity estimates. In **Chapter 6** we argue that the prediction interval is helpful in this context, because it shows the range of true treatment effects that is expected in future studies. Confidence intervals only relate to the mean treatment effect. In case of heterogeneity, prediction intervals will show a wider range of expected treatment effects than confidence intervals, and thus may lead to different conclusions. In roughly 70% of the statistically significant (random effects $p<0.05$) meta-analyses in the CDSR Issues 2009-2013, the 95% prediction interval suggested that the intervention effect could even be in the opposite direction, and in approximately 20% of these meta-analyses the opposite effect was as least as large as the summary estimate of the meta-analysis (e.g. the prediction interval contained an OR of 0.5 when the point estimate was an

OR of 2). Although we were not the first to describe the advantages of presenting prediction intervals, it is still far from common practice in meta-analyses of interventions.

Conclusion and discussion

The landscape of evidence based medicine is changing and the methodology of meta-analysis is maturing. This is described in **Chapter 7.** Meta-analyses are no longer restricted to the paired comparison of two interventions: network meta-analysis enables the comparison of a group of interventions for the same medical condition. Further there are many initiatives stimulating the data sharing of individual participant data. The availability of these data will allow a consistent statistical approach as well as more thorough insight in clinical reasons for the heterogeneity. Evaluating the variation in the strength of the effect across studies is a key feature of meta-analyses. Acceptance of this variation is crucial and fundamental to the use of a random-effects approach. Heterogeneity plays an important role, both in the statistical estimation of the summary effect and in the clinical interpretation of the results of a meta-analysis.

Nederlandse samenvatting

Het thema van het Cochrane Colloquium 2015 was “Filtering the information overload for better health decisions”, wat zoveel betekent als: de overvloed aan informatie filteren om betere beslissingen met betrekking tot de gezondheid te kunnen nemen. Per dag verschijnen er meer dan 75 publicaties over gerandomiseerde gecontroleerde klinische studies: een (nog immer) groeiende berg aan informatie. Een van de belangrijkste manieren om informatie uit deze informatieberg te destilleren is meta-analyse.

Reproduceerbaarheid van studieresultaten is een tweede issue dat hoog op de wetenschapsagenda staat. Resultaten blijken moeilijk te reproduceren, zelfs als het gaat om studies die gepubliceerd werden in tijdschriften met hoge impactcijfers. Publicatiebias en beperkte steekproefomvang van studies (met als gevolg random variatie) worden vaak als mogelijke oorzaken genoemd. Ook klinisch inhoudelijke of statistische verschillen tussen de studies zijn een mogelijke oorzaak voor de variatie in de resultaten. Deze variatie in resultaten van verschillende studies wordt tussen-studie heterogeniteit genoemd.

De belangrijkste doelstelling van dit proefschrift is inzicht geven in de methodologie voor meta-analyse, en dan met name in de rol die heterogeniteit tussen de studieresultaten hierin speelt. De standaardmethodologie is gebaseerd op benaderingen die valide zijn voor grote steekproeven, dat wil zeggen dat deze methodologie ervan uitgaat dat er voldoende studies beschikbaar zijn. In de praktijk zijn de meeste meta-analyses echter gebaseerd op slechts een paar studies die soms ook nog klein van omvang zijn.

De UK Cochrane Editorial Unit was zo vriendelijk ons de statistische gegevens van alle systematische reviews van klinische interventie studies uit de Cochrane Database of Systematic Reviews (CDSR) van 2009 tot 2013 ter beschikking te stellen. Deze gegevens hebben we gebruikt om verschillende aspecten van de meta-analyse methodologie empirisch te evalueren, en om realistische simulatiestudies te kunnen uitvoeren om zodoende nog meer inzicht te verkrijgen.

Meta-analyses van dierstudies – handleiding

De meeste medische meta-analyses zijn gericht op klinisch (humaan) onderzoek. Sommige vragen kunnen echter alleen beantwoord worden door experimentele dierstudies. Het aantal meta-analyses van dierstudies is groeiende, maar toch nog steeds veel lager dan het aantal meta-analyses van klinische studies, ondanks het feit dat de statistische methodologie vergelijkbaar is. Wellicht zijn de voordelen van een meta-analyse van dierstudies zelfs groter dan die van klinische studies. Resultaten van dierstudies zijn meer heterogeen dan die van humane studies, omdat dierstudies vaak meer variatie kennen wat betreft de onderzoekspopulatie (denk aan diersoorten), studieopzet, en andere studiekenmerken. Daarom kan het gecombineerde resultaat van een aantal studies (met behulp van meta-analyse) betrouwbaarder zijn dan de resultaten van een nieuwe studie. Daarnaast geldt dat als een meta-analyse overtuigende resultaten laat zien, er geen nieuwe studie gedaan hoeft te worden, waardoor er minder dieren hoeven te sterven. Bovendien maken meta-analyses op basis van dierstudies de tekortkomingen in de rapportage van deze studies zichtbaar. Daarom hopen we dat ook de rapportagekwaliteit zal toenemen met het aantal meta-analyses van dierstudies.

Om de methodologie uit te leggen en om dieronderzoekers te stimuleren meer meta-analyses uit te voeren, hebben we een inleidende handleiding geschreven, die te vinden is in **Hoofdstuk 2**. Dierstudies zijn echter (statistisch gezien) vaak complexer dan humane studies: ze kunnen erg klein zijn, er is veel heterogeniteit tussen de studies, ze gebruiken een keur aan verschillende soorten studie-opzet en de rapportage laat vaak te wensen over. Er zijn dus onderwerpen genoeg voor een uitgebreidere handleiding met meer geavanceerde statistische methoden. Maar hiervoor is eerst additioneel methodologisch onderzoek nodig.

Het gecombineerde resultaat van een meta-analyse op basis van een paar kleine studies is betrouwbaarder dan het resultaat van één grote studie

In **Hoofdstuk 3** vergelijken we twee benaderingen om een behandeling te evalueren: het uitvoeren van één nieuwe, grote studie versus het samenvatten van de resultaten van een aantal al bestaande studies door middel van een meta-analyse. We hebben dit geëvalueerd in samenhang met drie complicerende factoren. Ten eerste: aantal en omvang van de bestaande studies. Maakt het uit

voor het antwoord hoeveel studies er zijn, en hoe groot deze studies zijn? Ten tweede hebben we gekeken naar de invloed van publicatiebias: het verschijnsel dat studies met positieve resultaten vaker worden gepubliceerd dan studies met negatieve resultaten. Gevolg hiervan is dat het aantal beschikbare studies met positieve resultaten onevenredig groot is. De derde factor waar we rekening mee hebben gehouden is de mate van heterogeniteit tussen de studies.

Met behulp van uitgebreide simulaties hebben we aangetoond dat een meta-analyse een betere inschatting van de werkzaamheid van een behandeling gaf dan een nieuwe studie, met name als er sprake was van heterogeniteit tussen de studies, hetgeen het geval is in ongeveer 50% van de meta-analyses. Vanwege de heterogeniteit kunnen de resultaten van één enkele studie positiever of negatiever uitvallen dan gemiddeld het geval zou zijn. Maar als je slechts de resultaten van één studie kent, is het moeilijk inschatten hoe ver de gevonden resultaten afstaan van het effect dat je zou vinden als je beschikking had over voldoende studies. Simulatiestudies met een behandeling die gemiddeld geen effect had maar wel heterogeniteit, resulteerden in veel te hoge percentages fout positieve bevindingen, namelijk tot meer dan tien maal verhoogd, ook als de heterogeniteit maar laag was. Dit terwijl de foutpercentages correct waren voor meta-analyses op basis van twee of meer studies. Als er sprake was van publicatiebias waren de foutpercentages van de gecombineerde resultaten ook verhoogd, maar minder sterk dan voor de resultaten van één enkele studie. Zelfs als er slechts twee of drie studies beschikbaar waren voor de meta-analyse en als deze ook nog onderhevig waren aan publicatiebias en heterogeniteit, resulteerde de meta-analyse in betere, lagere percentages fout positieve bevindingen dan een nieuwe grote studie.

De Hartung-Knapp-Sidik-Jonkman methode werkt beter dan de standaard DerSimonian-Laird methode

In een random-effect meta-analyse beïnvloedt de grootte van de geschatte tussen-studie heterogeniteit hoe zwaar de primaire studies meewegen in het gecombineerde resultaat van de meta-analyse. De schatting van de heterogeniteit heeft dus direct invloed op het eindresultaat. Maar als de meta-analyse op slechts een paar studies is gebaseerd zal deze schatting niet erg nauwkeurig zijn, en dit komt vaak voor. De meta-analyses uit de CDSR 2009-2013 waren op maar weinig primaire studies gebaseerd: mediaan 4,

interkwartielafstand 2-6 studies. Toch is de standaardmethode voor random-effect meta-analyse nog steeds die van DerSimonian en Laird (DL), zelfs al is deze gebaseerd op de aanname dat er voldoende studies beschikbaar zijn, en ook al negeert deze methode de onnauwkeurigheid in de schatting van de heterogeniteit. Het is bekend dat de methode zoals beschreven door Hartung, Knapp en door Sidik en Jonkman (HKSJ) beter werkt, maar tot nu toe was er geen bewijs daarvoor in realistische situaties, waarbij studies sterk van grootte kunnen verschillen. In **Hoofdstuk 4** hebben we beide methoden vergeleken voor situaties waarin studies van verschillende groottes worden gecombineerd. We hebben simulaties uitgevoerd van meta-analyses op basis van 2–20 studies van verschillende grootte en met verschillende heterogeniteit. De resultaten laten zien dat de HKSJ methode resulteerde in lagere, meer adequate percentages fout positieve resultaten dan de DL methode. We hebben ook beide methoden toegepast op 689 meta-analyses van de CDSR jaargang 2012: het aantal "positieve" bevindingen (statistisch significant met een p-waarde <0.05) met de HKSJ benadering was veel lager dan met de DL benadering. Daarnaast hebben we aangetoond hoe DL resultaten omgezet kunnen worden in HKSJ resultaten door middel van een paar stappen, of door gebruik te maken van een door ons bijgeleverde Excel sheet.

Kleine studies waren meer heterogeen dan grote studies

Resultaten van kleine studies worden vaak geassocieerd met "kleine-studie-effecten": het verschijnsel dat de resultaten van kleine studies vaak wat extremer zijn dan die van grotere studies. Dit kan veroorzaakt worden door bias in de rapportage, maar ook door kwaliteitsissues, en beide komen vaker voor bij kleine studies. In **Hoofdstuk 5** hebben we onderzocht of er ook verschil is tussen kleine en grote studies in de mate van tussen-studie heterogeniteit. Hiervoor hebben we gebruikt gemaakt van 3263 meta-analyses van de CDSR 2009-2013. We hebben laten zien dat de schatting van de heterogeniteit tussen kleine studies gemiddeld groter was dan voor grote studies. Bovendien was ook de onnauwkeurigheid van deze schattingen voor de kleine studies groter. Resultaten van kleine studies verschillen meer van elkaar dan die van grote studies, zelfs na correctie voor de extra random variatie die veroorzaakt wordt door de kleinere steekproeven in de kleine studies.

Een pleidooi om predictie-intervallen in meta-analyses altijd te rapporteren

De mate van heterogeniteit tussen de studies bepaalt niet alleen hoe zwaar individuele studies meewegen in de meta-analyse, maar bevat ook klinisch relevante informatie. Als er heterogeniteit is (>0) impliceert dit namelijk dat het effect van een behandeling verschilt van studie tot studie. Dit komt in ongeveer de helft van de meta-analyses voor. Het betekent dat een behandeling in sommige situaties effectiever zal zijn dan in andere situaties. Dit is belangrijk voor de klinische praktijk, maar de meeste lezers van een publicatie over een meta-analyse weten niet goed hoe ze de schatting van de heterogeniteit klinisch moeten interpreteren. In **Hoofdstuk 6** beargumenteren we dat het rapporteren van een predictie-interval hierbij zou kunnen helpen, omdat dit aangeeft welke ware effecten er verwacht kunnen worden in een toekomstige studie. Predictie-intervallen zijn niet hetzelfde als betrouwbaarheidsintervallen. Betrouwbaarheidsintervallen geven aan hoe groot het geschatte gemiddelde behandelingseffect is, maar niet hoe groot de variatie hiervan is over de studies. Als er heterogeniteit is, zullen predictie-intervallen een breder bereik van verwachte effecten laten zien dan betrouwbaarheids-intervallen, en daarom ook tot andere conclusies kunnen leiden. In grofweg 70% van de statistisch significante (random effect p-waarde<0.05) meta-analyses in de CDSR 2009-2013 suggereerde het 95% predictie-interval dat het effect van de behandeling in sommige situaties nihil of zelfs tegengesteld kon zijn aan het gemiddelde effect, en in ongeveer 20% van deze meta-analyses was dit tegengestelde effect minstens zo groot als het gemiddelde effect van de meta-analyse (bijvoorbeeld, een predictie-interval dat een odds ratio van 0.5 bevat, terwijl de puntschatting een odds ratio van 2 betreft).

Weliswaar zijn wij niet de eersten die de voordelen van predictie-intervallen beschrijven, maar de praktijk laat zien dat de meeste publicaties van meta-analyses nog geen predictie-intervallen bevatten.

Conclusie en discussie

Meer en meer moeten medische behandelingen aantoonbaar effectief zijn (Evidence Based Medicine). Meta-analyses zijn daarom steeds relevanter en de vragen die ze moeten beantwoorden zijn dan ook complexer dan vroeger. Dit fenomeen wordt beschreven in **Hoofdstuk 7**. Van oudsher waren meta-analyses beperkt tot de vergelijking van twee behandelingen, gebaseerd op de

geaggregeerde resultaten van een aantal studies. Netwerk meta-analyse maakt het mogelijk om in één keer een vergelijking te maken tussen een groter aantal behandelingen voor dezelfde medische indicatie. Daarnaast zijn er steeds meer initiatieven die stimuleren dat de (geanonimiseerde) oorspronkelijke gegevens van individuele patiënten ook voor anderen dan de aanvankelijke onderzoekers beschikbaar komen. Als gebruik wordt gemaakt van individuele patiëntgegevens kan er voor alle studies in de meta-analyse een consistente statistische benadering worden toegepast, hetgeen de statistische heterogeniteit vermindert. Bovendien geven de individuele gegevens meer inzicht in de klinische oorzaken voor heterogeniteit dan de geaggregeerde gegevens.

Meta-analyse maakt het mogelijk om variatie in de sterkte van de werking van een behandeling over studies heen te bestuderen. Vaak krijgt deze variatie echter niet de aandacht die het verdient. Om een random-effect benadering goed te kunnen toepassen is het cruciaal dat deze variatie wordt geaccepteerd en bestudeerd. Heterogeniteit speelt een belangrijke rol, zowel bij het statistisch proces van het tot stand komen van het gecombineerde effect als in de klinische interpretatie van de resultaten van een meta-analyse.